2018

Developments in MCMC diagnostics and sparse Bayesian learning models

Anand Ulhas Dixit
Iowa State University

Developments in MCMC diagnostics and sparse Bayesian learning models

by

Anand Ulhas Dixit

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:
Vivekananda Roy, Major Professor
Alicia Carriquiry
Somak Dutta
Mark Kaiser
Dan Nettleton

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2018

DEDICATION

To Ramakrishna Paramhansa, Swami Vivekananda, Mata Sarada Devi and Shri Brahmachaitanya Gondavalekar Maharaj.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS

I would first like to thank my adviser Prof. Vivekananda Roy for his encouragement, support and valuable feedback throughout my PhD program. He has played a crucial role in my academic progress. I would also like to thank Prof. Dan Nettleton who showed confidence in my abilities and provided me a Research Assistant position at Plant Science Institute for two semesters and also served on my POS committee. I am also thankful to Prof. Alicia Carriquiry and Prof. Mark Kaiser for admitting me in the PhD program in August, 2013 and serving on my POS committee. I would also like to thank Prof. Somak Dutta for serving on my POS committee and always being available to provide me feedback. During my PhD program, Prof. Max Morris and Prof. Ken Koehler were Department chair at different time periods and they both equally supported and motivated me in my PhD journey. I would also like to thank Prof. Jarad Niemi for several insightful discussions in the advanced Bayesian methods course that I found to be beneficial for my academic progress.

My PhD journey would not have been possible without the unconditional love and support provided by my parents (Ulhas Dixit and Vaijayanti Dixit) and my sister (Vaidehi Dixit). Our family friend Dr. Rana Biswas and his family who reside in Ames have supported and encouraged me throughout my PhD program. I am also thankful to my friends Nehemias, Danny, Nick, Manju, Nate, MaryKate, Chirag, Elizabeth, Akshay, Satvik, Akshit, Anuj, Oscar, Andrew and Pritam for their academic and personal support throughout the PhD program.

During my time at the Department of Statistics, University of Mumbai I had the good fortune of learning from Prof. R. G. Shenoy, Prof. Smita Nabar, Prof. Meena Satam, Prof.

Shilpa Khare, Prof. Annapurna, Prof. Rajwadkar, Prof. Mangala Deshpande and Prof. Sushil Kulkarni. Their teachings enabled me to pursue my PhD degree. I would also like to thank my undergraduate teachers Prof. Pradnya Khandeparkar, Prof. Leela Subramanian, Prof. Kazi, Prof Hegde and Prof. Pallavi Rege for their guidance and motivation to purse higher degrees in Statistics. During my high school, I was lucky to have Prof. Smita Phadkar as my English teacher who played a crucial role in developing my english communication skills. I am indebted to her for her exemplary teaching. Last but not the least, thanks to Starbucks for providing an amazing atmosphere and drinks that enabled me to write my dissertation in an enjoyable space.

ABSTRACT

This dissertation consists of three research articles on the topic of Markov chain Monte Carlo (MCMC) diagnostics and sparse Bayesian learning models. The first article consists of MCMC diagnostic tools based on Kullback Leibler (KL) divergence and smoothing methods. These tools can assess the joint convergence of multiple variables and detect non convergence when MCMC chains get stuck at a particular mode of a multi modal stationary distribution. Further, in case of non convergence of multiple MCMC chains, the visualization tool can be used to investigate reasons for non convergence. The second article deals with assessing posterior propriety of some sparse Bayesian learning models. Relevance Vector Machine (RVM) is a popular sparse Bayesian learning model that assumes improper prior over its hyperparameters. We prove that this improper prior leads to an improper posterior. Further, we also provide necessary and sufficient conditions for posterior propriety of RVM. Additionally, we also prove the posterior impropriety of some Bayesian learning models that have a prior structure similar to that of RVM. In the third article, we propose to replace multiple penalties of RVM with a single penalty. The new model is named as single penalty relevance vector machine (SPRVM) and is analyzed using a semi Bayesian approach. The SPRVM allows for computation of Monte Carlo standard errors since we prove the geometric ergodicity of its associated Gibbs sampler. Such a Monte Carlo standard error cannot be computed in the case of RVM since the rate of convergence of its associated Gibbs sampler is not known. Thus, through these three articles we hope to make valuable additions to the literature of MCMC diagnostics and sparse Bayesian learning models.

CHAPTER 1. INTRODUCTION

Bayesian methods allow researchers to incorporate prior information in their analysis and express answers to the questions of interest in terms of probability statements. Therefore, they have been employed extensively in various disciplines. But there are methodological and theoretical challenges associated with Bayesian models. Consider the following Bayesian model,

$$y \sim f(y|\theta) \tag{1.1}$$

$$\theta \sim \pi(\theta), \tag{1.2}$$

where $f(y|\theta)$ is the data model and $\pi(\theta)$ is the prior density assumed on the parameter θ. For the above model, the posterior density of the parameter θ given the data y is given by,

$$\pi(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{m(y)}, \tag{1.3}$$

where $m(y)$ is the marginal likelihood, also known as the normalizing constant which is given by,

$$m(y) = \int_{\Theta} f(y|\theta)\pi(\theta)d\theta. \tag{1.4}$$

The posterior density given in (1.3) is often summarized by means of functions, posterior quantiles etc. which are of interest to researchers. The posterior mean of a function $g : \Theta \to \mathcal{R}^d$ is given by,

$$E(g(\theta)|y) = \int_{\Theta} g(\theta)\ \pi(\theta|y)d\theta. \tag{1.5}$$

Often, the marginal likelihood, i.e., the normalizing constant given in (1.4), is not known. Hence, the posterior density given in (1.3) and its mean given in (1.5) are not available in closed form. In such a scenario, one can produce a Monte Carlo estimate of the mean of

the posterior distribution if we can simulate observations from the posterior density given in (1.3). Given the intractability of the posterior distribution, it is often difficult to obtain independent and identically distributed (*iid*) observations, and hence one has to often rely on Markov chain Monte Carlo (MCMC) samplers to draw observations from an approximation to the posterior density given in (1.3). Now, MCMC samplers are iterative in nature and are expected to start producing observations from a close approximation to the posterior distribution as the Markov chain converges. In order to assess convergence of MCMC samplers, often empirical diagnostic tools are employed. Some of the empirical diagnostic tools currently available in the literature cannot assess the joint convergence of multiple variables and might falsely detect convergence when MCMC chains get stuck in a particular mode of a multi modal posterior distribution. We attempt to address these issues by proposing diagnostic tools based on Kullback Leibler divergence and smoothing methods in chapter 2 of this thesis.

In the absence of prior information, researchers often tend to assume improper priors, i.e., a function that is not a valid probability density function (for e.g. $\pi(\theta) \propto 1 \ \forall \theta \in \mathcal{R}$). With improper priors, posterior propriety is no longer guaranteed. Thus, in the case of an improper prior, we need to establish posterior propriety by bounding the normalizing constant, i.e., the marginal likelihood from above using a finite positive quantity. Posterior propriety cannot be checked using computational tools. In fact, Hobert and Casella (1996) show that MCMC draws corresponding to an improper posterior distribution did not provide any red flag to the user and seemed perfectly reasonable. Thus, posterior propriety can be established only through a theoretical exercise. Further, the conditions required for the lower bound of the normalizing constant to exist are necessary conditions for posterior propriety while the conditions required for the upper bound of the normalizing constant to exist are sufficient conditions for posterior propriety. Further, the improper priors that do not satisfy the necessary conditions lead to an improper posterior while the improper priors that satisfy

the sufficient conditions will lead to a proper posterior. Relevance vector machine (RVM) is a very popular sparse Bayesian learning model proposed by Tipping (2001) (cited more than 5000 times till date) that uses an improper prior. In chapter 3 of this thesis we establish both necessary and sufficient conditions for posterior propriety of RVM. Further, we show that the improper priors utilized by Tipping (2001) do not satisfy the necessary conditions and hence lead to an improper posterior. Additionally we also show that the sparse Bayesian models proposed by Mallick et al. (2005) and Figueiredo (2002) also lead to improper posteriors.

RVM proposed by Tipping (2001) involves multiple penalty parameters. In order to conduct a valid Bayesian analysis of RVM, one needs to use priors that satisfy the sufficient conditions established in chapter 3 of this thesis. After doing so, one can use RVM for prediction which is often based on the mean of the posterior predictive distribution. But, since the posterior predictive distribution is analytically intractable, its mean is also not available in closed form. In such a scenario, one can produce a Monte Carlo estimate of the mean of the posterior predictive distribution. If the Monte Carlo estimate is based on *iid* draws from the posterior predictive distribution, then one can easily estimate the corresponding standard error associated with the Monte Carlo estimate. But, in the case of RVM, it is not possible to obtain *iid* draws, and hence one has to often rely on MCMC draws. Since the MCMC draws are correlated by definition, producing an asymptotically valid standard error associated with the Monte Carlo estimate is challenging. In order to overcome this difficulty, the most common method is to establish geometric ergodicity, i.e., to show that the Markov chain used in the MCMC sampler converges to the stationary distribution at a geometric rate. Once geometric ergodicity has been established, one can compute a Monte Carlo standard error using standard methods like batch means and spectral variance (see Vats et al. (2018) and Vats et al. (2015)). In the case of RVM, often a Gibbs sampler is implemented whose rate of convergence is not known in the literature. Thus, in the case of RVM one cannot compute asymptotically valid standard errors associated with

the Monte Carlo estimate of the mean of the posterior predictive distribution. Therefore, in chapter 4 of the thesis, we propose single penalty relevance vector machine (SPRVM) model in which multiple penalty parameters are replaced by a single penalty parameter. In SPRVM, we establish the geometric ergodicity of the Gibbs sampler and hence can calculate both a Monte Carlo estimate of the mean of the posterior predictive distribution and its associated standard error. Thus, in chapter 4 we provide a single penalty approach to analyzing RVM that has both methodological as well as theoretical advantages over RVM. Some concluding remarks and future work suggestions are provided in chapter 5 of the thesis.

CHAPTER 2. MCMC DIAGNOSTICS FOR HIGHER DIMENSIONS USING KULLBACK LEIBLER DIVERGENCE

A paper published in the *Journal of Statistical Computation and Simulation*

Anand Dixit and Vivekananda Roy

Abstract

In order to simulate observations from an analytically intractable probability distribution (target distribution), researchers often utilize Markov Chain Monte Carlo (MCMC) samplers. Empirical diagnostic tools are commonly used to assess convergence of these MCMC samplers. In the existing literature of MCMC diagnostics, we have identified two areas for improvement. Firstly, the density based diagnostic tools currently available in the literature are not equipped to assess the joint convergence of multiple variables. Secondly, in case of multi-modal target distribution if the MCMC sampler gets stuck in one of the modes, then the current diagnostic tools may falsely detect convergence. In order to address these two issues, we propose two new diagnostic tools. The Tool 1 proposed in this article makes use of adaptive kernel density estimation, symmetric Kullback Leibler (KL) divergence and a testing of hypothesis framework to assess the joint convergence of multiple variables. In cases where Tool 1 detects divergence of multiple chains, started at distinct initial values, we propose a visualization tool that can help to investigate reasons behind their divergence. The Tool 2 proposed in this article makes a novel use of the target distribution (known up till the unknown normalizing constant), to detect divergence when an MCMC sampler gets stuck in one of the modes of a multi-modal target distribution. The usefulness of the tools

proposed in this article is illustrated using a multi-modal distribution, a mixture of bivariate normal distribution and a Bayesian logit model example.

2.1 Introduction

The process of simulating observations from a fully specified distribution is generally carried out using a traditional technique like inversion sampling. But if the distribution is analytically intractable, then there may not be any efficient methods for direct simulations from it, and in such cases, one often relies on Markov Chain Monte Carlo (MCMC) sampling techniques. Specifically, in Bayesian analysis, we often come across situations where the posterior distribution of parameters of interest is only known up to some unknown normalizing constant. In such situations, one often uses MCMC algorithms to produce approximate observations from the analytically intractable posterior distributions to make inference about the parameters of interest.

MCMC samplers are iterative in nature and require a starting observation. If $\pi(\theta)$ is the analytically intractable target distribution from which we wish to simulate observations, then MCMC samplers like the Gibbs sampler and the Metropolis Hasting sampler construct a Markov chain $\{\theta_n : n = 0, 1, 2, \cdots\}$ started at θ_0 such that the stationary distribution of the chain is equal to $\pi(\theta)$. Thus, as $n \to \infty$, the distribution of θ_n converges to $\pi(\theta)$. In the Bayesian inference framework, the target distribution is the posterior distribution of the unknown parameters θ given the data y and is commonly denoted as $\pi(\theta|y)$. Hence if the stationary distribution exists and is unique, then at some point the MCMC sampler will start producing approximate observations from the target distribution. Generally, deep mathematical analysis is needed to establish quantitative convergence bounds for determining the sample size (n) required for the Markov chain to be sufficiently close to the target distribution (see e.g. Rosenthal (1995), Jones and Hobert (2001), Roy and Hobert (2007)).

In the absence of such theoretical analysis, often empirical diagnostic tools are used to check the convergence of MCMC samplers.

In the early 90's there was an interesting debate on whether one should use multiple chains or a single long chain to diagnose convergence. Gelman and Rubin (1992) and Brooks and Gelman (1998) advocated the usage of multiple chains, while Raftery and Lewis (1992) and Geweke (1992) believed a single long chain is sufficient for assessing convergence. Gelman and Rubin (1992) used the potential scale reduction factor (PSRF) to monitor convergence in the univariate case. Suppose we are working with m chains and each chain has n iterations. Let $\{\theta_{ij} : i = 1, 2, \cdots, m \; and \; j = 1, 2, \cdots, n\}$ be the observations generated from the m chains. Then the PSRF, denoted by $\hat{R}$ is defined as,

$$\hat{R} = \frac{\hat{V}}{W}, \tag{2.1}$$

where

$\hat{V} = ((n-1)/n)W + (1 + (1/m))(B/n)$ is the pooled variance estimate,

$B/n = \sum_{i=1}^{m} (\bar{\theta_{i\cdot}} - \bar{\theta_{\cdot\cdot}})^2/(m-1)$ is the between chain variance estimate,

$W = \sum_{i=1}^{m} \sum_{j=1}^{n} (\theta_{ij} - \bar{\theta_{i\cdot}})^2/(m(n-1))$ is the within chain variance estimate and

$\bar{\theta_{i\cdot}}$ and $\bar{\theta_{\cdot\cdot}}$ are the i^{th} chain mean and the overall mean respectively where i = 1, 2, $\cdots$, m.

Brooks and Gelman (1998) came up with the multivariate PSRF (MPSRF) to diagnose convergence in the multivariate case. It is denoted by $\hat{R}_p$ and is given by,

$$\hat{R}_p = \max_a \frac{a^T \widehat{V^*} a}{a^T W^* a} = \frac{n-1}{n} + \left(1 + \frac{1}{m}\right) \lambda_1, \tag{2.2}$$

where $\widehat{V^*}$ is the pooled covariance matrix, W^* is the within chain covariance matrix, B^* is the between chain covariance matrix and λ_1 is the largest eigenvalue of the matrix $(W^{*-1}B^*)/n$. In this diagnostic tool, convergence is detected when $\hat{R} \approx 1$ in the univariate case and $\hat{R}_p \approx 1$ in the multivariate case. Raftery and Lewis (1992) proposed a univariate diagnostic tool based on the specified level of accuracy desired by the user in quantile estimation. In this

tool, the chain obtained from the sampler is used to find the number of initial iterations that should be discarded (burn-in period), how long the chain should be run after the burn-in and how should the chain be thinned in order to obtain the desired level of accuracy in quantile estimation. Geweke (1992) also proposed a univariate diagnostic tool in which he used a test statistic, to compare the mean of a function of the samples from two non overlapping parts of the chain. Usually, the choice is the first 10% of the chain and the last 50% of the chain. Thus, Gelman and Rubin (1992), Brooks and Gelman (1998), Raftery and Lewis (1992) and Geweke (1992) are all moment based diagnostic tools.

More recently, researchers have come up with density based diagnostic tools. Boone et al. (2014) and Hjorth and Vadeby (2005) used divergence measures to come up with univariate diagnostic tools. Boone et al. (2014) estimated the Hellinger distance between the kernel density estimates of two chains or two parts of a single chain. If the estimated distance was close to zero (i.e. less than 0.10), then the Markov chains were said to have converged else not. Hjorth and Vadeby (2005) used a measure imitating KL divergence to compare the empirical distributions of subsequences of chains to the empirical distribution of the whole chain, and in the case of multiple chains, they compared empirical distribution of individual chains to the empirical distribution of the combination of all chains.

The density based diagnostic tools mentioned before are univariate tools and cannot assess the convergence of multiple variables jointly. The Tool 1 proposed in this article, computes the adaptive kernel density estimate of the joint distribution of each multivariate chain, and then compares the estimated symmetric KL divergence between them to a cut-off value, to assess convergence. Since the adaptive kernel density estimation suffers from the curse of dimensionality, for higher dimensions, Tool 1 monitors convergence marginally i.e. one variable at a time, and since we determine the cut-off values for KL divergence measure using a testing of hypothesis framework, they can be easily adjusted for multiple comparison. Thus, Tool 1 is a density based diagnostic tool that can assess convergence of

multiple variables jointly. Other notable differences between Tool 1 and the density based diagnostic tools mentioned before are, Boone et al. (2014) used numerical integration to compute the estimated Hellinger distance and Hjorth and Vadeby (2005) compute differences between empirical distribution functions over a partition of the real line, while we provide a Monte Carlo estimate of the symmetric KL divergence to compare the adaptive kernel density estimates of multiple chains.

If the target distribution is multi-modal, then the MCMC chain might get stuck in one of the modes. In such cases, even if the MCMC sampler is run for a reasonably long time, it continues to produce observations around that mode. Many of the current diagnostic tools only make use of the iterations obtained from the MCMC samplers to diagnose convergence, and so, in cases where the chain gets stuck in one of the modes, they get fooled into thinking that the target distribution is unimodal, and hence, falsely detect convergence. Most of the times, we do not know a priori how many modes the target distribution has, and hence, even if we use multiple chains, there is a chance that all chains might be stuck at the same mode. In order to overcome this difficulty, we propose Tool 2 that makes a novel use of the target distribution known up till the unknown normalizing constant in the diagnostic tool. Yu (1994) proposed a tool which also incorporates the target distribution, in which she estimated the unknown normalizing constant using the MCMC samples, and then estimated the L^1 distance between the kernel density estimate of the chain and the estimated target distribution over a compact set, where the difference between the two was most likely. But if the MCMC sampler is stuck in a particular mode, then the normalizing constant estimator proposed by Yu (1994) is no longer reliable.

Many other diagnostic tools are available in the literature, and a very nice review of them can be found in Brooks and Roberts (1998). Our tools can be used when either a single chain or multiple chain samplers are available. This article is structured as follows. In Section 2.2, we provide definition and certain properties of KL divergence. In Section

2.3, we propose two new MCMC convergence diagnostic tools and a visualization tool. In Section 2.4, we provide three examples to illustrate the usefulness of the proposed diagnostic and visualization tools. Some concluding remarks are given in Section 2.5.

2.2 Kullback Leibler Divergence

KL divergence is a measure used to calculate the difference between two probability distributions. If $P(\theta)$ and $Q(\theta)$ are any two probability density functions on $\Theta \subseteq \mathbb{R}^d$, then the KL divergence between $P(\theta)$ and $Q(\theta)$ is defined as,

$$KL(P|Q) = \int_{\Theta} log\left(\frac{P(\theta)}{Q(\theta)}\right) P(\theta)d\theta. \tag{2.3}$$

Some important properties of the KL Divergence are as follows,

- $KL(P|Q) \geq 0$,
- $KL(P|Q) = 0$ iff $P = Q$ almost everywhere wrt the Lebesgue measure, and
- KL divergence is not symmetric in P and Q.

The KL divergence is not symmetric because $KL(P|Q)$ is the expected difference between *log* of densities P and Q with respect to P while $KL(Q|P)$ is the expected difference between *log* of densities Q and P with respect to Q. The symmetric KL divergence between P and Q, denoted by $KL_{sy}(P, Q)$ is given as,

$$KL_{sy}(P, Q) = \frac{KL(P|Q) + KL(Q|P)}{2}. \tag{2.4}$$

2.3 Diagnostic Tools

2.3.1 Tool 1

Let $\pi(\theta)$ be the target distribution where $\theta \in \Theta \subseteq \mathbb{R}^d$. In order to explore the target distribution, two chains are initiated at different starting points and each chain produces n

observations. As prescribed in Gelman and Rubin (1992), the starting points should be over dispersed with respect to the target distribution. Let $\{\theta_{ij} : i = 1, 2 \ \ and \ \ j = 1, 2, ..., n\}$ be the n observations obtained from each of the two chains where $\theta_{ij} \in \Theta \subseteq \mathbb{R}^d \ \forall \ i \ and \ \forall \ j$. The adaptive kernel density estimates of observations obtained from the two chains are denoted by $P_{1n}(\theta)$ and $P_{2n}(\theta)$ and are found by substituting $i = 1$ and $i = 2$ respectively in the following equation,

$$P_{in}(\theta) = \frac{1}{n} \sum_{j=1}^{n} \prod_{k=1}^{d} \frac{1}{h_j^{(k)}} \ K\left(\frac{\theta^{(k)} - \theta_{ij}^{(k)}}{h_j^{(k)}}\right), \tag{2.5}$$

where,

$\theta_{ij}^{(k)}$ denotes the k^{th} dimension in the j^{th} observation of the i^{th} chain, where $i = 1, 2$; $j = 1, 2, \cdots, n$ and $k = 1, 2, \cdots, d$,

$\theta^{(k)}$ denotes the k^{th} dimension of a d dimensional vector at which the adaptive kernel density estimate is evaluated,

$\{h_j^{(k)} : j = 1, 2, \cdots, n \ and \ k = 1, 2, \cdots, d\}$ are smoothing parameters and

$K(\cdot)$ is a Gaussian kernel.

In (2.5), the smoothing parameters are chosen using Silverman (1986) (Sec 5.3.1) wherein observations in sparse regions are assigned Gaussian kernels with high bandwidth and observations in high probability regions are assigned Gaussian kernels with low bandwidth. In our examples, we use the kepdf function in the R package pdfCluster (Azzalini and Menardi (2014)) to compute the adaptive kernel density estimate of the chains. The KL divergence between P_{1n} and P_{2n}, denoted by $KL(P_{1n}|P_{2n})$ and KL divergence between P_{2n} and P_{1n}, denoted by $KL(P_{2n}|P_{1n})$ can be obtained after substituting appropriate values of i and j in the equation given below,

$$KL(P_{in}|P_{jn}) = \int_{\Theta} \Big(log\big(P_{in}(\theta)\big) - log\big(P_{jn}(\theta)\big) \Big) \ P_{in}(\theta) d\theta. \tag{2.6}$$

The symmetric KL divergence between P_{1n} and P_{2n}, denoted by $KL_{sy}(P_{1n}, P_{2n})$ is given below,

$$KL_{sy}(P_{1n}, P_{2n}) = \frac{KL(P_{1n}|P_{2n}) + KL(P_{2n}|P_{1n})}{2}. \tag{2.7}$$

We can find the Monte Carlo estimate of $KL(P_{1n}|P_{2n})$ and $KL(P_{2n}|P_{1n})$ using (2.8) and (2.9) respectively,

$$\widehat{KL}(P_{1n}|P_{2n}) = \frac{1}{n}\sum_{i=1}^{n}\Big\{log\big(P_{1n}(\theta^*_{1i})\big) - log\big(P_{2n}(\theta^*_{1i})\big)\Big\}, \tag{2.8}$$

where $\{\theta^*_{1i}\}_{i=1}^n$ are the observations simulated from $P_{1n}(\theta)$ using the technique proposed by Silverman (1986)(Sec 6.4.1). Similarly,

$$\widehat{KL}(P_{2n}|P_{1n}) = \frac{1}{n}\sum_{i=1}^{n}\Big\{log\big(P_{2n}(\theta^*_{2i})\big) - log\big(P_{1n}(\theta^*_{2i})\big)\Big\}, \tag{2.9}$$

where $\{\theta^*_{2i}\}_{i=1}^n$ are the observations simulated from $P_{2n}(\theta)$ using the technique proposed by Silverman (1986)(Sec 6.4.1). An estimate of $KL_{sy}(P_{1n}, P_{2n})$ is then given below,

$$\widehat{KL}_{sy}(P_{1n}, P_{2n}) = \frac{\widehat{KL}(P_{1n}|P_{2n}) + \widehat{KL}(P_{2n}|P_{1n})}{2}. \tag{2.10}$$

Adaptive kernel density estimation suffers from the curse of dimensionality. Hence we need to increase our sample size (n) as the dimension (d) increases in order to obtain a good estimate of the symmetric KL divergence between any two distributions.

In order to find the appropriate sample size (n) required for achieving convergence when univariate and bivariate chains are drawn from similar distributions, we conduct a simulation study. In the univariate case, for each n, we generate 1000 datasets each from $f_1 \equiv N(0,1)$ and $f_2 \equiv N(0,1)$. Let $\hat{f}_1$ and $\hat{f}_2$ be the adaptive kernel density estimates of observations drawn from f_1 and f_2 respectively. The estimated symmetric KL divergence between f_1 and f_2 for each pair can be computed using (2.10) in which P_{1n} and P_{2n} are replaced by $\hat{f}_1$ and $\hat{f}_2$. The true symmetric KL divergence between f_1 and f_2 is known to be zero. Thus we can then find the bias, standard deviation and root mean square error (RMSE) of $\widehat{KL}_{sy}(\hat{f}_1, \hat{f}_2)$.

Table 2.1: Bias, Standard Deviation and RMSE of (a) $\widehat{KL}_{sy}(\hat{f}_1, \hat{f}_2)$ where $\hat{f}_1$ and $\hat{f}_2$ are adaptive kernel density estimates of observations drawn from $f_1 \equiv N(0,1)$ and $f_2 \equiv N(0,1)$ (b) $\widehat{KL}_{sy}(\hat{f}_3, \hat{f}_4)$ where $\hat{f}_3$ and $\hat{f}_4$ are adaptive kernel density estimates of observations drawn from $f_3 \equiv N(\mathbf{0}, I_2)$ and $f_4 \equiv N(\mathbf{0}, I_2)$.

(a) Univariate Distribution				**(b) Bivariate Distribution**			
n	Bias	SD	RMSE	n	Bias	SD	RMSE
1000	0.0066	0.0041	0.0078	3000	0.0113	0.0029	0.0116
1500	0.0046	0.0026	0.0053	6000	0.0070	0.0016	0.0072
2000	0.0039	0.0021	0.0045	9000	0.0053	0.0011	0.0054
2500	0.0032	0.0018	0.0036	12000	0.0043	0.0009	0.0044
3000	0.0027	0.0015	0.0031	15000	0.0037	0.0007	0.0038

In the bivariate case, for each n, we generate 1000 datasets each from $f_3 \equiv N(\mathbf{0}, I_2)$ and $f_4 \equiv N(\mathbf{0}, I_2)$ and carry out a similar procedure as before to find the bias, standard deviation and RMSE of $\widehat{KL}_{sy}(\hat{f}_3, \hat{f}_4)$. The results are tabulated in Table 2.1.

In Table 2.1 we observe that for both univariate and bivariate chains, the bias, standard deviation and RMSE go on reducing as the sample size (n) increases. In order to assess the convergence of Markov chains, we intend to accurately estimate the symmetric KL divergence up till two decimal points. In Table 2.1, we observe that for $n = 2000$ in the univariate case and for $n = 12,000$ in the bivariate case, the bias, standard deviation and RMSE are significantly small, and hence, the symmetric KL divergence can be estimated efficiently up till two decimal points. Thus if we wish to use Tool 1 for assessing convergence, we need to run the chains for at least 2000 iterations in the univariate case and for at least 12,000 iterations in the bivariate case. We next check if the above mentioned sample sizes hold true, when instead of the Gaussian distribution the above simulation study is carried out using a heavy tailed distribution, skewed distribution or a distribution with dependent coordinates. For this purpose, the above simulation study was repeated with t distribution

(df=5), chi-square distribution (df=10) and bivariate normal distribution (with correlation coefficient equal to 0.3). The sample sizes prescribed above were found to be sufficient for efficiently estimating the symmetric KL divergence up till two decimal points in these cases. The tabulated results are provided in the Appendix. Thus we can safely use the Gaussian distribution for studying the bias, standard deviation and RMSE associated with the symmetric KL divergence estimator.

Tool 1 will detect convergence i.e. indicate that the chains have mixed adequately, when the estimated symmetric KL divergence between P_{1n} and P_{2n} will be close to zero. Hence we now need to identify cut-off points, so that estimated symmetric KL divergence below or equal to that point will indicate convergence. Boone et al. (2014) carried out a simulation study and came up with a criteria, wherein if the estimated Hellinger distance between the kernel density estimates of univariate chains is less than 0.10, then the Markov chains have converged else not.

We will utilize a testing of hypothesis framework to come up with cut-off points. In our framework, the null hypothesis states that the Markov chains have diverged, i.e., the chains have not yet mixed adequately and our alternative hypothesis states that the Markov chains have converged, i.e., the chains have mixed adequately. In this scenario, the probability of Type 1 error will be the probability of concluding that the Markov chains have converged when in fact they have not. We would like to limit this probability to some level α which is typically chosen to be 0.05. As mentioned earlier, we will be estimating the symmetric KL divergence up till two decimal points, hence the cut-off value should also be reported up till the second decimal point. In order to find our cut-off values, we will first generate 1000 datasets each from two dissimilar distributions and the maximum value of C for which $P(\widehat{KL}_{sy}(P_{1n}, P_{2n}) \leq C) \leq \alpha = 0.05$, i.e. the probability of Type 1 error is less than 0.05, will be our cut-off value.

In the univariate case, we will generate 1000 datasets with suitable sample size (n) each from $f_5 \equiv N(0,1)$ and $f_6 \equiv N(\mu,1)$ where $\mu \neq 0$. The adaptive kernel density estimate of observations drawn from f_5 and f_6 are denoted by $\hat{f}_5$ and $\hat{f}_6$ respectively. The maximum value of C for which $P(\widehat{KL}_{sy}(\hat{f}_5, \hat{f}_6) \leq C) \leq \alpha = 0.05$ will be our cut-off value in the univariate case. In the bivariate case, we will generate 1000 datasets with a suitable sample size (n) each from $f_7 \equiv N(\mathbf{0}, I_2)$ and $f_8 \equiv N(\mu \mathbb{1}_2, I_2)$ where $\mu \neq 0$ and then carry out a similar procedure as before to identify our cut-off value in the bivariate case. By limiting the probability of Type 1 error, we are exercising control over variability of estimated symmetric KL divergence, but we need to make sure that its bias does not affect the cut-off value adversely, and hence for this purpose, the suitable sample size (n) should be chosen in such a way that the bias is significantly small and does not affect the first two decimal points of the estimated value. The μ can be chosen by the users as per their requirement. If the users are dealing with a very sensitive experiment, then the choice of μ should be small otherwise μ can be chosen to be slightly bigger. For simulation studies corresponding to identifying cut-off values using dissimilar distributions, Boone et al. (2014) utilized $N(0,1)$ and $N(0.2835,1)$ since the true Hellinger distance between them is known to be 0.10. Hence a possible choice for μ can be 0.2835. For this choice of μ, as mentioned before, the sample size (n) should be chosen based on the bias of $\widehat{KL}_{sy}(\hat{f}_5, \hat{f}_6)$ and $\widehat{KL}_{sy}(\hat{f}_7, \hat{f}_8)$ for univariate and bivariate cases respectively.

The symmetric KL divergence between any two univariate Gaussian distributions with the same variance parameter, and any two multivariate Gaussian distributions with the same covariance matrix can be computed analytically using (2.11) and (2.12),

$$KL_{sy}(g_1, g_2) = \frac{(\mu_1 - \mu_2)^2}{2\sigma^2}, \tag{2.11}$$

where $g_1 \equiv N(\mu_1, \sigma^2)$ and $g_2 \equiv N(\mu_2, \sigma^2)$, $\mu_1 \in R$, $\mu_2 \in R$, and

$$KL_{sy}(g_3, g_4) = \frac{1}{2}\Big\{(\mu_4 - \mu_3)^T \Sigma^{-1} (\mu_4 - \mu_3)\Big\}, \tag{2.12}$$

Table 2.2: Bias, Standard Deviation and RMSE of (a) $\widehat{KL}_{sy}(\hat{f}_5, \hat{f}_6)$ where $\hat{f}_5$ and $\hat{f}_6$ are adaptive kernel density estimates of observations drawn from $f_5 \equiv N(0,1)$ and $f_6 \equiv N(0.2835, 1)$ (b) $\widehat{KL}_{sy}(\hat{f}_7, \hat{f}_8)$ where $\hat{f}_7$ and $\hat{f}_8$ are adaptive kernel density estimates of observations drawn from $f_7 \equiv N(\mathbf{0}\ \mathbb{1}_2, I_2)$ and $f_8 \equiv N(\mathbf{0.2835}\ \mathbb{1}_2, I_2)$.

(a) Univariate Distribution				**(b) Bivariate Distribution**			
n	Bias	SD	RMSE	n	Bias	SD	RMSE
1000	0.0054	0.0145	0.0154	3000	0.0079	0.0119	0.0143
1500	0.0035	0.0116	0.0122	6000	0.0037	0.0084	0.0092
2000	0.0024	0.0098	0.0101	9000	0.0029	0.0071	0.0077
2500	0.0019	0.0090	0.0092	12000	0.0022	0.0059	0.0063
3000	0.0018	0.0081	0.0083	15000	0.0016	0.0051	0.0053

where $g_3 \equiv N(\mu_3, \Sigma)$ and $g_4 \equiv N(\mu_4, \Sigma),\ \mu_3 \in R^d,\ \mu_4 \in R^d$.

Thus using (2.11) and (2.12), for $\mu = 0.2835$, true symmetric KL divergence between f_5 and f_6 was found to be 0.04 and the true symmetric KL divergence between f_7 and f_8 was found to be 0.08. We then compute the bias, standard deviation and RMSE of $\widehat{KL}_{sy}(\hat{f}_5, \hat{f}_6)$ and $\widehat{KL}_{sy}(\hat{f}_7, \hat{f}_8)$. The results are provided in Table 2.2.

In Table 2.2 we observe that, for $n = 2000$ in the univariate case and $n = 12,000$ in the bivariate case, the bias associated with $\widehat{KL}_{sy}(\hat{f}_5, \hat{f}_6)$ and $\widehat{KL}_{sy}(\hat{f}_7, \hat{f}_8)$ is significantly small and does not affect the first two decimal points of the estimate. In Table 2.2 we also observe that, the standard deviation and RMSE are also reasonably small for $n = 2000$ and $n = 12,000$ in the univariate and bivariate cases respectively. Hence we can carry out our cut-off procedure for $\mu = 0.2835$ with $n = 2000$ in the univariate case and $n = 12,000$ in the bivariate case. The Type 1 error associated with different cut-off values is given in Table 2.3.

In Table 2.3 we observe that $C = 0.02$ and $C = 0.06$ are ideal cut-off points for univariate and bivariate distributions respectively. The users should also be aware that if they

Table 2.3: Probability of Type 1 error associated with different cut-off values in the univariate and bivariate case.

(a) Univariate Distribution		**(b) Bivariate Distribution**	
C	$P(\widehat{KL}_{sy}(\hat{f}_5, \hat{f}_6) \leq C)$	C	$P(\widehat{KL}_{sy}(\hat{f}_7, \hat{f}_8) \leq C)$
0.01	0.001	0.06	0.001
0.02	0.040	0.07	0.097
0.03	0.242	0.08	0.687

choose a larger sample size than the one prescribed before, then the Type 1 error associated with the cut-off value will further reduce. Thus, if the user chooses μ to be 0.2835 and at $n = n^*$, estimated symmetric KL divergence (rounded to two decimal points) is found to be less than or equal to the cut-off value, then Tool 1 indicates that the chains have mixed adequately at that point, hence the user should then discard the first n^* observations and iterations obtained thereafter will be considered as approximate observations from the target distribution. On the other hand, if the estimated symmetric KL divergence (rounded to two decimal points) is found to be strictly greater than the cut-off value, then Tool 1 indicates that chains have not yet mixed adequately and the user should run the chain longer. If found necessary, the user can thin the chain and consider an initial burn-in before applying Tool 1.

Tool 1 can be used for more than two dimensions as well. For a multivariate chain, we recommend assessing convergence marginally, i.e. one variable at a time. We found our cut-off values using a testing of hypothesis approach, wherein our level of significance was chosen to be $\alpha = 0.05$. In the case of multiple comparison, we will adjust our level of significance using Bonferroni's correction, so that the overall type 1 error does not go beyond $\alpha = 0.05$. Hence, if the Markov chain is d dimensional (where $d > 2$), then our level of significance for each comparison will be α/d. Using this adjusted level of significance in

our cut-off procedure, we obtain an appropriate cut-off point for multiple comparison. For example, if our Markov chain is 10 dimensional, then in Table 2.3 we observe that $C = 0.01$ is an ideal cut-off point, since it maintains a type 1 error of less than $\alpha/10 = 0.005$ for each comparison. This ability to maintain the overall type 1 error at α, by adjusting the cut-off value in the case of multiple comparison, is another advantage of our tool over the tools proposed by Boone et al. (2014). Applying Tool 1 in the univariate case is similar to the tool proposed by Boone et al. (2014). Hence, in case of multi-modal target distribution, if both chains are stuck at the same mode, then this tool is also prone to failure.

In case of multiple chains, say m chains, we will find the estimated symmetric KL divergence between each of the $\binom{m}{2}$ combination of chains and find the maximum among them. If the maximum estimated symmetric KL divergence is less than or equal to the cut-off value, then Tool 1 indicates that the chains have converged else not. If the user wishes to use a single chain, then one can estimate the symmetric KL divergence between the adaptive kernel density estimate of any two non overlapping parts of the chain.

In cases where the state space is bounded, adaptive kernel density estimation might suffer from boundary bias if high probability regions are closer to the boundary. But the objective of Tool 1 is to check if the two chains have mixed adequately or not, hence, even if it is found that the adaptive kernel density estimate of the two chains suffers from boundary bias, the final objective is not affected as long as the same density estimation procedure is used for both chains. The user must also note that, if the sample simulated from the adaptive kernel density estimate of the chain contain several observations from outside the state space, and if the target distribution is not expected to have a lot of mass close to the boundary, then it is an indication that chains have not captured the target distribution adequately and thus Tool 1 indicates divergence in such a situation.

To implement Tool 1 for two univariate chains with $n = 2,000$ and two bivariate chains with $n = 12,000$, it takes approximately 3.81s and 109.86s respectively on an Intel (R)

Core (TM) i5-6300U 2.40GHz machine running Windows 10. For multivariate chains, we implement Tool 1 marginally i.e. one variable at a time, which can be done in parallel.

2.3.2 Visualization Tool

Suppose a user is using multiple chains, say m chains (where $m \geq 3$). Further suppose that application of Tool 1 revealed that the m chains have not mixed adequately and thus chains have not yet converged. This indication of divergence could be due to a variety of reasons. A common reason for divergence is formation of clusters among multiple chains. A visualization tool can be helpful for identifying these clusters.

Peltonen et al. (2009) had proposed a visualization tool based on Linear Discriminant Analysis and Discriminant Component Analysis which can be used to complement the diagnostic tools proposed by Gelman and Rubin (1992) and Brooks and Gelman (1998). Similarly the visualization tool described in this section will complement Tool 1 proposed in Subsection 2.3.1.

In this tool we utilize the tile plot. As mentioned before, in case of multiple chains, say m chains, Tool 1 will find the estimated symmetric KL divergence between each of the $\binom{m}{2}$ combinations and report the maximum among them. In the visualization tool, we will utilize the individual values provided by estimated symmetric KL divergence between each of the $\binom{m}{2}$ distinct combinations. If the estimated symmetric KL divergence for a particular combination is less than or equal to the cut-off value, then we will utilize a "Grey" tile to represent that the two chains belong to the same cluster else we will use a "black" tile to represent that the two chains belong to different clusters.

In the case of a multivariate chain, we monitor convergence marginally i.e. one variable at a time. Hence two multivariate chains will be considered to be in the same cluster, only if the estimated symmetric KL divergence for each variable is less than or equal to the cut-off value, which has been adjusted for multiple comparison. For further investigation of a

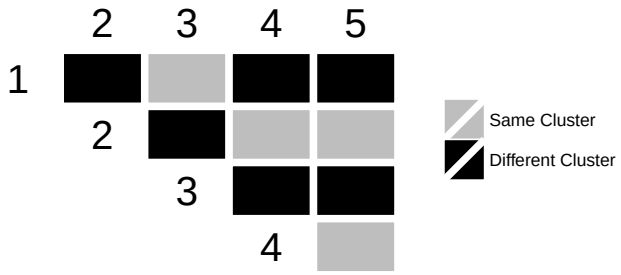

Figure 2.1: Application of the visualization tool in which chain 1 and chain 3 are drawn from N(0, 1) while chain 2, chain 4 and chain 5 are drawn from N(10, 1).

multivariate Markov chain, the user can consider the following steps.

Consider a d dimensional Markov chain initialized at m different points. Suppose these m chains (where $m \geq 3$), were grouped into q clusters. The visualization tool is utilized when Tool 1 indicates divergence i.e. $2 \leq q \leq m$. In cases where Tool 1 indicates divergence, for further investigation, the user can choose a chain from each cluster and implement the visualization tool marginally i.e. one variable at a time. This will help the user identify, which among the d variables are responsible for inadequate mixing among the m multivariate chains.

In order to provide an illustration of the visualization tool, suppose we run 5 chains for 5,000 iterations each wherein the 1^{st} and the 3^{rd} chain are drawn from $N(0, 1)$ while the 2^{nd}, 4^{th} and 5^{th} chain are drawn from $N(10, 1)$. The application of the visualization tool for these five chains is provided in Figure 2.1. As expected, Figure 2.1 indicates presence of two clusters wherein chain 1 and chain 3 form a cluster while chain 2, chain 4 and chain 5 form another cluster.

2.3.3 Tool 2

Suppose the target density is as follows,

$$\pi(\theta) = \frac{g(\theta)}{k},\ \theta \in \Theta, \tag{2.13}$$

where k is the unknown normalizing constant.

Suppose a single Markov chain is run for n iterations and the observations obtained are $\{\theta_{1j}\}_{j=1}^{n}$. Let $P_{1n}(\theta)$ denote the adaptive kernel density estimate of the observations as mentioned in (2.5). The KL divergence between $P_{1n}(\theta)$ and $\pi(\theta)$ is given below,

$$KL(P_{1n}|\pi) = G_n + log\ k, \tag{2.14}$$

where

$G_n = \int_{\Theta} log\big(P_{1n}(\theta)\big)\ P_{1n}(\theta)d\theta - \int_{\Theta} log\big(g(\theta)\big)\ P_{1n}(\theta)d\theta.$

In the implementation of Tool 2, we will assume that $KL(P_{1n}|\pi) \to 0$ as $n \to \infty$. Under this assumption, $G_n \to -log(k)$ as $n \to \infty$. Hence $\exp(-G_n) \to k$ as $n \to \infty$. Thus, an estimator of the normalizing constant based on KL divergence between $P_{1n}(\theta)$ and $\pi(\theta)$, denoted by $\hat{k}$ is given below,

$$\hat{k} = \exp\left(-\frac{1}{n}\sum_{i=1}^{n}\left\{ log\big(P_{1n}(\theta_{1i}^{*})\big) - log\big(g(\theta_{1i}^{*})\big)\right\}\right), \tag{2.15}$$

where $\{\theta_{1i}^{*}\}_{i=1}^{n}$ are the observations simulated from $P_{1n}(\theta)$ using the technique proposed by Silverman (1986)(Sec 6.4.1).

If the chain has converged, then the quantity T_2 given below will be close to 0,

$$T_2 = \frac{|\ \hat{k} - k\ |}{k}. \tag{2.16}$$

But T_2 contains the normalizing constant (k), which is unknown. Now, the unknown normalizing constant (k) is nothing but the integral given below,

$$k = \int_{\Theta} g(\theta)\ d\theta. \tag{2.17}$$

Table 2.4: Finding appropriate sample size (n) required for implementation of Tool 2 when the samples are indeed from the target distribution i.e. standard normal distribution.

(a) Univariate Distribution		**(b) Bivariate Distribution**	
n	$P(T_2^* \leq 0.05)$	n	$P(T_2^* \leq 0.05)$
100	0.631	1000	0.390
500	0.871	3000	0.658
1000	0.950	6000	0.869
2000	0.981	9000	0.955
3000	0.997	12000	0.985

Hahn (2005) came up with Cuba library which provides implementation of general purpose multidimensional integration algorithm. The R package R2Cuba (Bouvier and Kiu (2015)) provides the implementation of Cuba library in R. Using the "divonne" function in R2Cuba (Bouvier and Kiu (2015)) we will evaluate the integral given in (2.17) and thus obtain an estimate of the normalizing constant based on numerical integration which we will denote as k^*. The user can also use the "adaptIntegrate" function in the R package cubature (Johnson and Narasimhan (2013)) to produce an estimate k^* of the unknown normalizing constant (k). Replacing k by k^* in (2.16) we obtain T_2^* given below,

$$T_2^* = \frac{\mid \hat{k} - k^* \mid}{k^*}. \tag{2.18}$$

In the case of multi-modal target distribution, if the Markov chain gets stuck in one of the modes, then the adaptive kernel density estimate of the chain will estimate the density around that particular mode really well, while completely ignoring the rest of the density. The estimate $\hat{k}$ is a Monte Carlo estimate of the unknown normalizing constant (k), based on the KL divergence between the adaptive kernel density estimate of the chain and the target distribution. Thus, the evaluation of $\hat{k}$ involves drawing n observations from the adaptive kernel density of the chain. But, since the adaptive kernel density estimate cap-

tures only one mode, the estimate $\hat{k}$ only records the KL divergence between the adaptive kernel density estimate of the chain and the target distribution for the state space around that particular mode. Hence $\hat{k}$ will likely underestimate the true quantity k, if the Markov chain gets stuck in one of the modes. On the other hand, k^* will provide a good estimate of the unknown normalizing constant (k) since it uses numerical integration to integrate over the entire state space. Thus T_2^* can be interpreted as the percentage of the target distribution not yet captured by the Markov chain. A Markov chain that captures at least 95% of the target distribution can be considered to be producing approximate observations from the target distribution. Using this interpretation of T_2^*, we came up with a cut-off value of 0.05 wherein if $T_2^* > 0.05$, then Tool 2 indicates that the Markov chain has not yet captured the target distribution adequately.

As seen earlier, Tool 2 assumes (i) $\exp(-G_n) \to k$ as $n \to \infty$, (ii) $\hat{k}$ is a consistent Monte Carlo estimate of $\exp(-G_n)$ and (iii) k^* is an estimate of k based on multidimensional numerical integration. Since (i) and (ii) depend on the sample size (n), it is important to know their convergence rates as a function of n. But currently in the literature, a theoretical proof of the convergence of the adaptive kernel density estimate based on a Markov chain to its stationary distribution, with respect to the KL divergence is not available. In the absence of such a theoretical result, it is difficult to find the convergence rates of (i) and (ii) as a function of n. Hence we conduct a simulation study to choose a suitable sample size (n). Since the cut-off value is based on the interpretation of T_2^*, we will conduct a slightly different simulation study to get an intuition of the sample size (n) required to implement Tool 2. In the univariate case we will consider $f_1 \equiv N(0,1)$ to be our target distribution, generate 1000 datasets of sample size (n) each from $f_1 \equiv N(0,1)$ and find $\hat{k}$ using (2.15) while k^* will be obtained by numerically integrating the kernel of f_1, thus we can then find T_2^* for each dataset using (2.18). Then we estimate $P(T_2^* \leq 0.05)$ i.e. probability of concluding convergence when the chain is indeed from the target distribution. The sample size (n) for

which the estimate of $P(T_2^* \leq 0.05)$ is significantly high, will be our prescribed sample size (n) for Tool 2. In the bivariate case, similar procedure is carried out for $f_3 \equiv N(\mathbf{0}, I_2)$. The results are given in Table 2.4.

In Table 2.4 we observe that for $n = 2000$ in the univariate case and $n = 12,000$ in the bivariate case, the estimated probability of concluding that the Markov chain has converged, given the fact that it is indeed drawn from the target distribution is really high. This provides us an indication that $n = 2000$ and $n = 12,000$ are sufficient for implementing Tool 2 in univariate and bivariate distributions respectively. We also check if the prescribed sample sizes are sufficient when the target distribution is heavy tailed, skewed or is a distribution with dependent coordinates and hence we replicated the above study with t distribution (df=5), chi-square distribution (df=10) and bivariate normal distribution (with correlation coefficient equal to 0.3). We observed that the sample sizes mentioned above are sufficient. The detailed results are provided in the Appendix.

Tool 2 is specifically designed for detecting divergence when the target distribution is multi-modal and the chain gets stuck in one of the modes. Thus, if the user observes that even after running the chain for a reasonably long time (sample size prescribed above), the value of T_2^* was found to be greater than 0.05, then it is highly likely that the chain is stuck in one of the modes and has not yet traveled through the whole state space. The value of T_2^* will also tell the user the percentage of the target distribution not yet traveled by the Markov chain.

Tool 2 involves the target distribution, hence if the target distribution is bounded and has a lot of mass close to the boundary, then Tool 2 will be affected by boundary bias. Due to boundary bias, the sample generated from the adaptive kernel density estimator might contain observations from outside the state space and *log* of the target distribution without the unknown normalizing constant (i.e. $log(g(\theta))$) is not defined at these points. A possible solution to this situation is to consider a bootstrap sample instead of sampling from the

adaptive kernel density estimate. Also, in the case where the bounded target distribution has very little mass close to the boundary and the sample generated from the adaptive kernel density estimate contains none or negligible number of observations from outside the state space, we can simply ignore those and implement Tool 2.

Users must be aware that Tool 2 is vulnerable to poor adaptive kernel density estimation. Further, estimating the unknown normalizing constant using numerical integration is challenging, hence we do not claim that Tool 2 will solve all problems related to diagnosing convergence of Markov chains in the case of multi-modal target distributions. But, we hope that the tool will help the users understand the challenges associated with it, and thus further boost research in this direction.

To implement Tool 2 for a univariate chain with $n = 2,000$ and a bivariate chain with $n = 12,000$, it takes approximately 1.31s and 35.68s respectively on an Intel (R) Core (TM) i5-6300U 2.40GHz machine running Windows 10.

2.4 Examples

2.4.1 Six Mode Example

This example was proposed by Leman et al. (2009). Suppose the target density is as follows,

$$\pi(x, y) \propto \exp\left(\frac{-x^2}{2}\right) \exp\left(\frac{((csc\ y)^5 - x)^2}{2}\right), \tag{2.19}$$

where $-10 \leq x, y \leq 10$.

The contour plot of the target distribution known up to the normalizing constant is given in Figure 2.2 and the marginal distribution of X and Y known up to the normalizing constant is given in Figure 2.3. The visualization of joint and marginal distribution clearly show that the target distribution is multi-modal in nature.

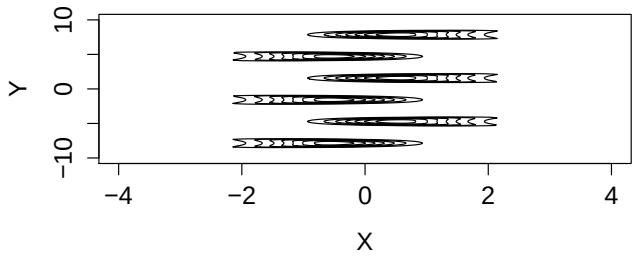

Figure 2.2: Contour plot of the target distribution in the Six Mode Example.

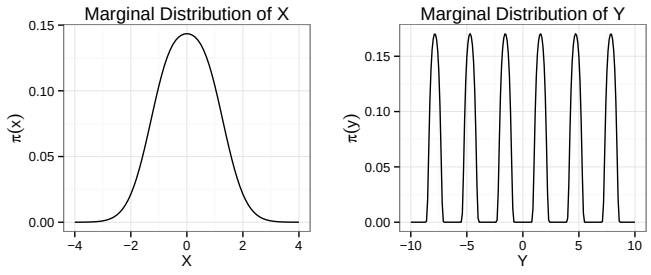

Figure 2.3: Marginal Distribution of X and Y in the six mode example.

We will consider two different cases to illustrate the application of the diagnostic tools and the visualization tool proposed in Section 2.3. In order to draw observations from the target distribution, we will use a Metropolis within Gibbs sampler in which X is drawn first and then Y.

Case 1

In this case, we will run four chains wherein two chains (chain 1 and chain 2) will be started at a particular mode while the remaining two chains (chain 3 and chain 4) will be started at some other mode. Each of the four chains were run for $n = 30,000$ iterations. The adaptive kernel density estimates of the four chains are visualized in Figure 2.4.

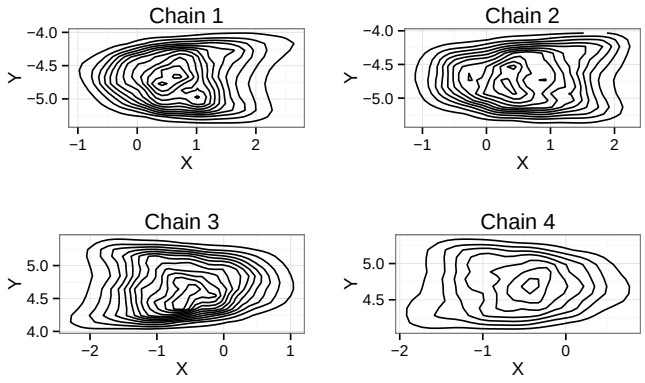

Figure 2.4: Visualizations of the adaptive kernel density estimates of the four chains in case 1.

In order to assess the convergence of the above Markov chains, several diagnostic tools were implemented and the results are presented in Table 4.1. In Table 4.1 we observe that the PSRF proposed by Gelman and Rubin (1992), MPSRF proposed by Brooks and Gelman (1998), Hellinger distance approach proposed by Boone et al. (2014) and Tool 1 proposed in Section 2.3 correctly indicate that the chains have not yet converged.

Table 2.5: Application of various MCMC convergence diagnostic tools for Case 1 in the Six Mode Example.

Variable	$\hat{\mathbf{R}}$	**H Dist**	$\hat{\mathbf{R}}_\mathbf{p}$	**Biv KL Tool 1**
X	1.63	0.60	21.39	104.89
Y	23.80	1.00		

Tool 1 has correctly identified that Markov chains have diverged, now to identify clusters among those four chains, we will utilize our visualization tool proposed in Section 2.3. The result of the visualization tool given in Figure 2.5 shows that there are two clusters among

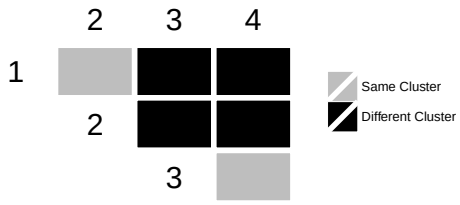

Figure 2.5: Application of the Visualization Tool in the case 1 of the Six Mode example.

four chains wherein chain 1 and chain 2 form a cluster, while chain 3 and chain 4 form another.

Case 2

In this case as well we will run four chains but all the chains will be started at the same mode. All the four chains were run for $n = 30,000$ iterations and the adaptive kernel density estimates of the four chains are visualized in Figure 2.6.

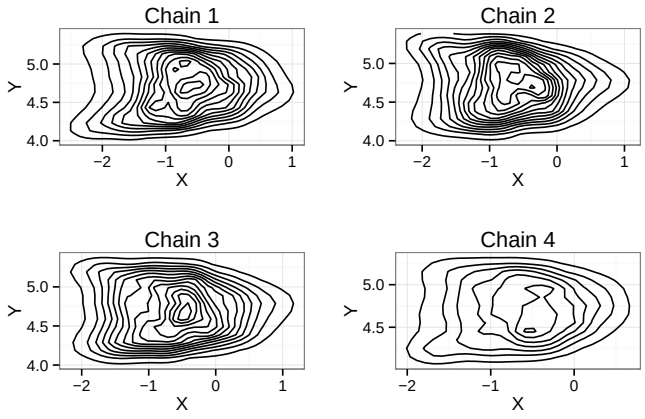

Figure 2.6: Visualizations of the adaptive kernel density estimates of the four chains in case 2.

Convergence diagnostic tools used in case 1 were applied in this case as well and the results obtained are tabulated in Table 2.6. In Table 2.6, we observe that since all chains are

stuck at the same mode, the PSRF, the MPSRF and the Hellinger distance approach falsely detect convergence. Now, Tool 2 requires only one chain and since the PSRF, MPSRF and the Hellinger distance suggest that the four chains are similar, we can simply choose any one among them. Now, $T_2^* = 0.88$ is significantly greater than zero and thus indicates that the chains are stuck at the same mode, further it also indicates that 88% of the target distribution is not yet captured by the Markov chain. Thus Tool 2 is both computationally cheap as well as efficient in detecting divergence in the case of multi-modal target distributions.

Table 2.6: Application of various MCMC convergence diagnostic tools for Case 2 in the Six Mode Example.

Variable	$\hat{\mathbf{R}}$	**H Dist**	$\hat{\mathbf{R}}_\mathbf{p}$	$\mathbf{T}_\mathbf{2}^*$
X	1.00	0.09	1.00	0.88
Y	1.00	0.03		

2.4.2 Mixture of Bivariate Normal

Suppose the target density is as given below,

$$\pi\Big((x,y)^T\Big) = \frac{1}{2}\,\phi_2\Big((x,y)^T,\ 0,\ \Sigma_1\Big) + \frac{1}{2}\,\phi_2\Big((x,y)^T,\ 0,\ \Sigma_2\Big) \tag{2.20}$$

where $(x,y)^T \in R^2$, $\phi_2\Big((x,y)^T,\ 0,\ \Sigma_i\Big)$ is the density of bivariate normal distribution with mean 0, covariance matrix Σ_i, evaluated at $(x,y)^T$ and $\Sigma_i = \begin{bmatrix} 1 & \rho_i \\ \rho_i & 1 \end{bmatrix}$ for i = 1, 2.

The target distribution with $\rho_1 = 0.99$ and $\rho_2 = -0.99$ is plotted in Figure 2.7. To simulate observations from the above target distribution, we will utilize the MCMC function in the R package adaptMCMC (Scheidegger (2012)) which uses an adaptive Metropolis algorithm. We will use four chains wherein two chains each are started in neighboring corners. All four chains were run for $n = 1000$ iterations. The chains are visualized in

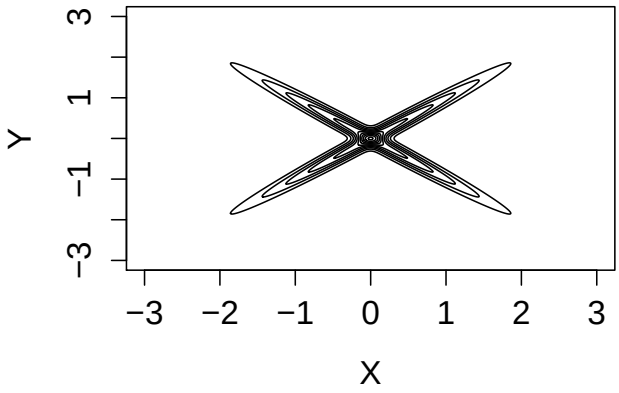

Figure 2.7: Joint distribution of the mixture of two bivariate normals in which the first component is highly positively correlated while the second component is highly negatively correlated.

Figure 2.8. In Figure 2.8 we observe that since each component is so highly correlated, chains get stuck in the component in which they were started. Hence the chains have clearly not mixed adequately.

In order to detect divergence in this example, one needs to assess the convergence of both X and Y jointly. Thus, marginal empirical convergence diagnostic tools like PSRF proposed by Gelman and Rubin (1992) and Hellinger distance approach proposed by Boone et al. (2014) are not applicable in such cases. The MPSRF has the ability to assess the joint convergence of X and Y, but it falsely detects convergence ($\hat{R}_p = 1.03$) since it averages the within chain covariances and thus the positive covariances and the negative covariances cancel each other out. Using the bivariate KL Tool 1, the maximum estimated symmetric KL divergence between chains was found to be 1.98 which is significantly greater than the cut-off of $C = 0.06$. The bivariate KL Tool 1 requires at least 12,000 observations to provide a good estimate hence simply comparing the estimated symmetric KL divergence to the cut-off value is not enough. Thus, we must also look at the probability of observing an estimated

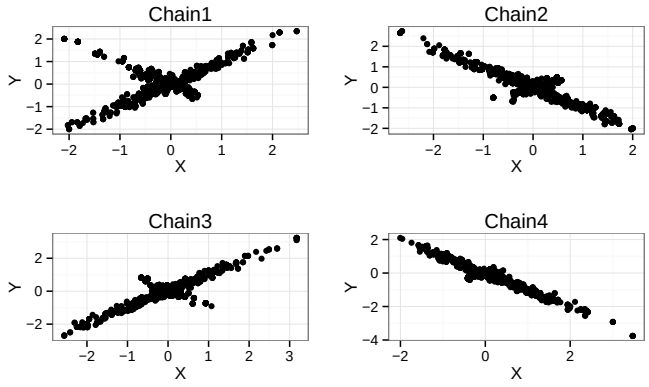

Figure 2.8: Visualizations of the four chains in the mixture of bivariate normals example in which $\rho_1 = -\rho_2 = 0.99$.

symmetric KL divergence of 1.98 or less when the chains with $n = 1000$ are drawn from different distributions i.e. $N(0\mathbb{1}_2, I_2)$ and $N(\mu\mathbb{1}_2, I_2)$ where $\mu = 0.2835$. This probability which can also be looked upon as the p-value in terms of the hypothesis framework given in Section 2.3 was found to be very large (i.e. greater than 0.999). Thus we do not reject our null hypothesis and concluded that the chains have not yet mixed adequately. Thus, in this example we observe that, even if the target distribution is unimodal, MPSRF proposed by Brooks and Gelman (1998) is vulnerable to false indication of convergence. If the above chains are run for a longer period, then all the chains travel through both the components and mix adequately.

2.4.3 Logit Model

We use the "Anguilla_train" dataset provided in the R package dismo (Hijmans et al. (2016)) to fit a Bayesian logit model with presence or absence of short finned eel considered as the response variable and the six other variables were included as covariates. The six covariates are: summer air temperature (SeqSumT), distance to coast (DSDist), area with indigenous forest (USNative), average slope in the upstream catchment (USSlope), maximum

downstream slope (DSMaxSlope) and Fishing Method (categorical variable with five classes namely electric, mixture, net, spot and trap). This example was also used by Boone et al. (2014) to illustrate the usefulness of the Hellinger distance approach. The model is as follows,

$$
\begin{aligned}
Y_i &\sim Bernoulli(\mu_i), \\
\mu_i &= \frac{exp(X_i^T\beta)}{1+exp(X_i^T\beta)}, \\
\beta &\sim N(\mathbf{0}, 100\ I_{10}).
\end{aligned}
$$

In order to estimate the parameters in the above model, three chains were run using the MCMClogit function in the R package MCMCpack (Martin et al. (2011)). We considered an initial burn-in of 30,000 as suggested by Boone et al. (2014). Since the dimension of the MCMC chain is large, we will apply Tool 1 marginally to diagnose the convergence of each of the parameter. After the initial burn-in, each of the three chains were run for $n = 3000$, $n = 15,000$ and $n = 30,000$ iterations and for each n, the convergence was diagnosed using PSRF, Hellinger distance and univariate KL Tool 1. The results are tabulated in Table 2.7.

Table 2.7: Application of various MCMC convergence diagnostic tools to the Bayesian logit model.

	$n = 3,000$			$n = 15,000$			$n = 30,000$		
Variable	$\hat{\mathbf{R}}$	**H Dist**	**Tool 1**	$\hat{\mathbf{R}}$	**H Dist**	**Tool 1**	$\hat{\mathbf{R}}$	**H Dist**	**Tool 1**
(Intercept)	1.01	0.12	0.06	1.00	0.07	0.02	1.00	0.06	0.01
SeqSumT	1.01	0.13	0.07	1.00	0.07	0.02	1.00	0.05	0.01
DSDist	1.00	0.13	0.05	1.00	0.06	0.01	1.00	0.05	0.01
USNative	1.00	0.14	0.08	1.00	0.07	0.02	1.00	0.06	0.01
M - mix	1.01	0.13	0.05	1.00	0.07	0.01	1.00	0.05	0.01
M - net	1.01	0.17	0.11	1.00	0.06	0.01	1.00	0.06	0.01
M - spot	1.01	0.13	0.08	1.00	0.09	0.02	1.00	0.06	0.01
M - trap	1.02	0.15	0.08	1.00	0.08	0.02	1.00	0.06	0.01
DSMaxSlope	1.00	0.09	0.03	1.00	0.06	0.01	1.00	0.05	0.01
USSlope	1.00	0.10	0.03	1.00	0.09	0.03	1.00	0.06	0.01

The Markov chain is 10 dimensional, hence in order to maintain an overall type 1 error rate of $\alpha = 0.05$, we will use Bonferroni's correction to adjust our cut-off point. For $C = 0.01$, the type 1 error for each comparison is less than $\alpha/10 = 0.005$ while the overall type 1 error is less than $\alpha = 0.05$. The PSRF ($\hat{R}$) indicates that the chains have mixed adequately at $n = 3000$ and hence iterations obtained from 3000 onwards can be used to compute parameters estimates. But Hellinger distance approach and multivariate KL Tool 1 indicate that chains have not yet converged and hence the user must run the sampler longer. The Hellinger distance approach indicates convergence of all parameters at $n = 15,000$ while the multivariate KL Tool 1 indicates convergence of all parameters at $n = 30,000$. Since the multivariate KL Tool 1 adjusts its cut-off point for multiple comparison, it is advisable for the user to use iterations from 30,000 onwards for making inference about the parameters of interest.

2.5 Conclusion

In this article, we have provided two new MCMC convergence diagnostic tools based on KL divergence and smoothing methods. The advantage of the first tool over existing MCMC convergence diagnostic tools is that, it has the ability to assess the joint convergence of multiple variables. For multivariate chains, we assess convergence marginally and recalibrate the cut-off point using Bonferroni's correction to maintain the overall type 1 error at α. Due to the use of Bonferroni's correction, Tool 1 can be conservative in the case of large number of variables. But in the case of MCMC diagnostics, a conservative tool is preferable as it provides the user greater assurance, that the chain is producing approximate observations from the target distribution, when it indicates convergence. In the case where the first tool indicates divergence of multiple MCMC chains, the user can use the visualization tool to further investigate reasons behind the divergence of multiple MCMC chains. The advantage

of the second tool over existing MCMC convergence diagnostic tools is that, it is equipped to detect divergence when MCMC chains get stuck in a particular mode of a multi-modal target distribution. Tool 2 is vulnerable if multidimensional numerical integration does not provide a good estimate of the unknown normalizing constant. Thus the proposed methods provide a useful addition to the set of available MCMC diagnostic tools and are equipped to detect non convergence of chains when other methods might fail to do so. A possible future study involves deriving a theoretical proof of convergence of the adaptive kernel density estimate based on Markov chain samples to its stationary distribution with respect to KL divergence measure.

2.6 Appendix: Results Corresponding to Additional Simulations

2.6.1 Additional Simulations for Tool 1

In Subsection 2.3.1, we have conducted a simulation study to find the sample size (n), required to estimate the symmetric KL divergence efficiently when the chains are drawn from similar univariate and bivariate Gaussian distributions. Here we will provide details about the study being replicated with t distribution (df = 5), chi-square distribution (df=10) and bivariate normal distribution (with correlation coefficient equal to 0.3).

Table 2.8: Bias, Standard Deviation and RMSE of (a) $\widehat{KL}_{sy}(\hat{h}_1, \hat{h}_2)$ where $\hat{h}_1$ and $\hat{h}_2$ are adaptive kernel density estimates of observations drawn from $h_1 \equiv t_5(0, 1)$ and $h_2 \equiv t_5(0, 1)$ (b) $\widehat{KL}_{sy}(\hat{h}_3, \hat{h}_4)$ where $\hat{h}_3$ and $\hat{h}_4$ are adaptive kernel density estimates of observations drawn from $h_3 \equiv t_5(\mathbf{0}, I_2)$ and $h_4 \equiv t_5(\mathbf{0}, I_2)$.

(a) Univariate t distribution			
n	Bias	SD	RMSE
1000	0.0074	0.0046	0.0088
1500	0.0053	0.0031	0.0061
2000	0.0043	0.0025	0.0049
2500	0.0035	0.0019	0.0040
3000	0.0030	0.0016	0.0034

(b) Bivariate t distribution			
n	Bias	SD	RMSE
3000	0.0111	0.0031	0.0115
6000	0.0068	0.0019	0.0070
9000	0.0051	0.0013	0.0053
12000	0.0041	0.0009	0.0042
15000	0.0035	0.0012	0.0038

In the case of univariate t distribution, we generate 1000 datasets of sample size (n) each from $h_1 \equiv t_5(0,1)$ and $h_2 \equiv t_5(0,1)$ and using the procedure discussed in Subsection 2.3.1, find the bias, standard deviation and RMSE of $\widehat{KL}_{sy}(\hat{h}_1, \hat{h}_2)$ where $\hat{h}_1$ and $\hat{h}_2$ are the adaptive kernel density estimates of h_1 and h_2 respectively. In the case of bivariate t distribution, the above study is performed with $h_3 \equiv t_5(\mathbf{0}, I_2)$ and $h_4 \equiv t_5(\mathbf{0}, I_2)$. The results are given in Table 2.8.

Table 2.9: Bias, Standard Deviation and RMSE of (a) $\widehat{KL}_{sy}(\hat{h}_5, \hat{h}_6)$ where $\hat{h}_5$ and $\hat{h}_6$ are adaptive kernel density estimates of observations drawn from $h_5 \equiv \chi^2_{10}$ and $h_6 \equiv \chi^2_{10}$ (b) $\widehat{KL}_{sy}(\hat{h}_7, \hat{h}_8)$ where $\hat{h}_7$ and $\hat{h}_8$ are adaptive kernel density estimates of observations drawn from $h_7 \equiv \chi^2_{10} \times \chi^2_{10}$ and $h_8 \equiv \chi^2_{10} \times \chi^2_{10}$.

(a) Univariate chi sq distribution			
n	Bias	SD	RMSE
1000	0.0058	0.0038	0.0070
1500	0.0044	0.0026	0.0051
2000	0.0036	0.0021	0.0042
2500	0.0030	0.0017	0.0034
3000	0.0026	0.0015	0.0029

(b) Bivariate chi sq distribution			
n	Bias	SD	RMSE
3000	0.0101	0.0027	0.0104
6000	0.0063	0.0015	0.0064
9000	0.0047	0.0011	0.0049
12000	0.0038	0.0009	0.0039
15000	0.0033	0.0007	0.0034

Table 2.10: Bias, Standard Deviation and RMSE of $\widehat{KL}_{sy}(\hat{h}_9, \hat{h}_{10})$ where $\hat{h}_9$ and $\hat{h}_{10}$ are adaptive kernel density estimates of observations drawn from $h_9 \equiv N(\mathbf{0}, \Sigma_1)$ and $h_{10} \equiv N(\mathbf{0}, \Sigma_2)$ where $\Sigma_1 = \Sigma_2 = [1, 0.3; 0.3, 1]$.

Biv dep normal distribution			
n	Bias	SD	RMSE
3000	0.0108	0.0028	0.0111
6000	0.0066	0.0016	0.0068
9000	0.0050	0.0011	0.0051
12000	0.0041	0.0009	0.0042
15000	0.0035	0.0007	0.0036

In the case of univariate chi-square distribution, we conduct a similar study as above with $h_5 \equiv \chi^2_{10}$ and $h_6 \equiv \chi^2_{10}$, while in the case of bivariate chi-square distribution we use $h_7 \equiv \chi^2_{10}$ x χ^2_{10} and $h_8 \equiv \chi^2_{10}$ x χ^2_{10}. The results are provided in Table 2.9. In the case of bivariate normal distribution (with correlation coefficient equal to 0.3), we perform the Tool

1 simulation study with $h_9 \equiv N(\mathbf{0}, \Sigma_1)$ and $h_{10} \equiv N(\mathbf{0}, \Sigma_2)$ where $\Sigma_1 = \Sigma_2 = [1, 0.3; 0.3, 1]$. The results are tabulated in Table 2.10.

In Table 2.8, 2.9 and 2.10 we observe that the bias, standard deviation and RMSE goes on reducing as the sample size (n) increases. Also in each of the three distributions we observe that for $n = 2000$ in the univariate case and $n = 12000$ in the bivariate case, we can produce a good estimate of the symmetric KL divergence since the bias, standard deviation and RMSE become reasonably small.

2.6.2 Additional Simulations for Tool 2

With respect to Tool 2, in Subsection 2.3.3, we conducted a simulation study wherein chains were drawn from the univariate and bivariate Gaussian distribution to find the sample size (n) required to implement Tool 2. Here we provide details about the study being implemented for t distribution (df $= 5$), chi-square distribution (df=10) and bivariate normal distribution (with correlation coefficient equal to 0.3).

Table 2.11: Finding appropriate sample size (n) required for implementation of Tool 2 when the chain is indeed from the target distribution i.e. t distribution with df=5 in this case.

(a) Univariate t distribution	
n	$P(T_2^* \leq 0.05)$
100	0.827
250	0.989
500	0.998
750	1.000
1000	1.000

(b) Bivariate t distribution	
n	$P(T_2^* \leq 0.05)$
100	0.347
500	0.918
1000	0.998
1500	1.000
2000	1.000

In the case of univariate t distribution, we generate 1000 datasets each of sample size (n) from $h_1 \equiv t_5(0, 1)$ and compute T_2^* for each dataset using the procedure discussed in Subsection 2.3.3. Then we can estimate $P(T_2^* \leq 0.05)$ for different values of sample size (n). In the case of bivariate t distribution, the study is performed for $h_3 \equiv t_5(\mathbf{0}, I_2)$. The results

are given in Table 2.11. In the case of univariate and bivariate chi-square distribution, we perform the above study for $h_5 \equiv \chi^2_{10}$ and for $h_7 \equiv \chi^2_{10}$ x χ^2_{10} respectively. The results are provided in Table 2.12. In the case of bivariate dependent normal distribution, the above study was carried out with $h_9 \equiv N(\mathbf{0}, \Sigma_1)$, where $\Sigma_1 = [1, 0.3; 0.3, 1]$. The results are given in Table 2.13.

Table 2.12: Finding appropriate sample size (n) required for implementation of Tool 2 when the chain is indeed from the target distribution i.e. chi-square distribution with df=10 in this case.

(a) Univariate chi sq distribution	
n	$P(T_2^* \leq 0.05)$
100	0.591
500	0.903
1000	0.979
2000	0.999
3000	1.000

(b) Bivariate chi sq distribution	
n	$P(T_2^* \leq 0.05)$
1000	0.417
3000	0.800
6000	0.986
9000	0.999
12000	1.000

Table 2.13: Finding appropriate sample size (n) required for implementation of Tool 2 when the chain is indeed from the target distribution i.e. bivariate dependent normal distribution (with correlation coefficient equal to 0.3) in this case.

Bivariate dep normal distribution	
n	$P(T_2^* \leq 0.05)$
1000	0.323
3000	0.517
6000	0.740
9000	0.882
12000	0.950

In Subsection 2.3.3, where the target distribution was taken to be Gaussian, we had observed that $n = 2000$ and $n = 12,000$ were found to be sufficient for implementing Tool 2 in the univariate and bivariate case respectively. In Table 2.11, 2.12 and 2.13 we observe that for the above mentioned sample sizes, estimate of $P(T_2^* \leq 0.05)$ is significantly

large. Thus even when the target distribution is heavy tailed, skewed or has dependent coordinates, the above mentioned sample sizes are sufficient for capturing the target distribution.

Bibliography

Azzalini, A. and Menardi, G. (2014). Clustering via nonparametric density estimation: The R package pdfCluster. *Journal of Statistical Software*, 57(11):1–26.

Boone, E., Merrick, J., and Krachey, M. (2014). A Hellinger distance approach to MCMC diagnostics. *Journal of Statistical Computation and Simulation*, 84(4):833–849.

Bouvier, A. and Kiu, K. (2015). *R2Cuba: Multidimensional Numerical Integration*. R package version 1.1-0.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7:434–455.

Brooks, S. P. and Roberts, G. O. (1998). Convergence assessment techniques for Markov chain Monte Carlo. *Statistics and Computing*, 8(4):319–335.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–472.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculations of posterior moments. *Bayesian Statistics 4*, J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, eds., Oxford University Press, Oxford:169–193.

Hahn, T. (2005). Cuba - a library for multidimensional numerical integration. *Computer Physics Communications*, 168(2):78 – 95.

Hijmans, R. J., Phillips, S., Leathwick, J., and Elith, J. (2016). *dismo: Species Distribution Modeling*. R package version 1.0-15.

Hjorth, U. and Vadeby, A. (2005). Subsample distribution distance and MCMC convergence. *Scandinavian Journal of Statistics*, 32:313–326.

Johnson, S. G. and Narasimhan, B. (2013). *cubature: Adaptive multivariate integration over hypercubes*. R package version 1.1-2.

Jones, G. and Hobert, J. P. (2001). Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Statist. Sci.*, 16(4):312–334.

Leman, S. C., Chen, Y., and Lavine, M. (2009). The multiset sampler. *Journal of the American Statistical Association*, 104(487):1029–1041.

Martin, A. D., Quinn, K. M., and Park, J. H. (2011). MCMCpack: Markov chain Monte Carlo in R. *Journal of Statistical Software*, 42(9):22.

Peltonen, J., Venna, J., and Kaski, S. (2009). Visualizations for assessing convergence and mixing of Markov chain Monte Carlo simulations. *Computational Statistics & Data Analysis*, 53(12):4453 – 4470.

Raftery, A. E. and Lewis, S. M. (1992). How many iterations in the Gibbs sampler. *Bayesian Statistics 4*, J.M. Bernardo, A.F.M. Smith, A.P. Dawid and J.O. Berger, eds., Oxford University Press, Oxford:763–773.

Rosenthal, J. S. (1995). Minorization conditions and convergence rates for Markov chain Monte Carlo. *Journal of the American Statistical Association*, 90(430):558–566.

Roy, V. and Hobert, J. P. (2007). Convergence rates and asymptotic standard errors for Markov chain Monte Carlo algorithms for Bayesian probit regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):607–623.

Scheidegger, A. (2012). *adaptMCMC: Implementation of a generic adaptive Monte Carlo Markov Chain sampler*. R package version 1.1.

Silverman, B. W. (1986). Density estimation for statistics and data analysis. *Chapman & Hall, London*.

Yu, B. (1994). Estimating the L1 error of kernal estimators based on Markov samplers. *Technical Report, UC Berkeley*.

CHAPTER 3. POSTERIOR IMPROPRIETY OF SOME SPARSE BAYESIAN LEARNING MODELS

A paper under review

Anand Dixit and Vivekananda Roy

Abstract

Modern datasets often have a significantly greater number of covariates than observations. For such datasets, often the objective is to predict the response variable for previously unobserved values of the covariates. Researchers have proposed some sparse Bayesian learning models that take a reproducing kernel Hilbert space (RKHS) approach to carry out the task of prediction. Among the class of sparse Bayesian learning models, relevance vector machines (RVM) is very popular. In this article we show that RVM and some other sparse Bayesian learning models with hyperparameter values currently used in the literature are based on improper posteriors. Further, we also provide necessary and sufficient conditions for posterior propriety of RVM.

3.1 Introduction

Advances in technology have resulted in huge growth of availability of high dimensional data in a variety of disciplines. For such datasets, often the number of covariates (p) is much greater than the number of observations (n), and the objective is to construct a model to predict the response variable for previously unobserved values of the covariates. As an example, such high dimensional datasets are abundant in the field of genetics, where

the number of subjects involved in the experiment is small, but for each subject, data corresponding to millions of markers are available. The objective in such datasets is to predict the physiological phenotype given the genomic information. If $p < n$, then one can fit a suitable linear model using a traditional statistical technique like ordinary least squares (OLS). But if $p > n$, then OLS is no longer applicable, and hence one can rely on penalized methods such as the least absolute shrinkage and selection operator (LASSO) (Tibshirani (1996)) or ridge regression (Hoerl and Kennard (1970)) to find a suitable model. But, both LASSO and ridge regression are penalized regression techniques that perform variable selection among the class of linear models. Hence, in case of $p > n$, if we wish to explore nonlinear class of models, we can estimate a function (f) from a functional space ($\mathcal{H}$) using the following Tikhonov regularization,

$$\min_{f \in \mathcal{H}} \left[\sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda ||f||^2_{\mathcal{H}} \right], \tag{3.1}$$

where $\{y_i, x_i\}_{i=1}^n$ is the training data such that $y_i \in \mathcal{R}$ $\forall i$ and $x_i \in \mathcal{R}^p$ $\forall i$, $L(\cdot, \cdot)$ is the loss function, λ is the penalty parameter, $\mathcal{H}$ is the functional space and $||\cdot||_{\mathcal{H}}$ is the norm defined on $\mathcal{H}$.

Since a functional space is infinite dimensional, the solution of (3.1) can also be infinite dimensional. Hence there is a possibility that we cannot use it for practical purposes. Wahba (1990) proved that, if the functional space is a reproducing kernel Hilbert space (RKHS), then the solution is finite dimensional and is given by

$$f(x) = \sum_{j=1}^{n} k(x, x_j)\beta_j, \tag{3.2}$$

where $k(\cdot, \cdot)$ is a reproducing kernel and $\{\beta_j\}_{j=1}^n$ are some unknown coefficients. The formal definition of RKHS and reproducing kernel can be found in Berlinet and Thomas-Agnan (2011).

Tipping (2001) used the above finite dimensional solution in a hierarchical Bayesian model to introduce the relevance vector machine (RVM) (see also Tipping (2000) and Bishop and

Tipping (2000)). The prior structure of RVM has been chosen in such a manner that it will produce a sparse solution and hence will lead to better predictions. RVM is a very popular sparse Bayesian learning model that is typically used for prediction (The paper Tipping (2001) has been cited more than 5000 times till date.).

In Bayesian analysis, in the absence of prior knowledge on the parameters, it is a common practice to use the so called 'non-informative' or 'objective' priors. A popular example of an objective prior is the Jeffreys's prior, which is considered to be non-informative in nature (Jeffreys (1961)). It is a function that is directly proportional to the square root of the determinant of the Fisher information matrix and hence can be computed easily in several cases. Often the objective priors turn out to be improper distributions i.e., functions that do not integrate to a finite number. For Bayesian models involving improper priors, the posterior distribution of the parameters given the data is not guaranteed to be proper. Hence, in such cases, it is necessary to show that the normalizing constant associated with the posterior distribution is bounded above by a finite constant otherwise there is a possibility that the posterior distribution is improper and drawing inference from an improper posterior distribution is equivalent to drawing inference from a function that integrates to infinity.

The priors imposed in RVM involve an improper prior on the hyperparameters. We prove that this improper prior leads to an improper posterior distribution. Additionally, we also derive necessary and sufficient conditions for the posterior propriety of RVM. The necessary conditions will help past researchers of RVM to check if the improper prior used by them leads to an improper posterior, and the sufficient conditions will provide guidelines for future researchers to choose prior distributions that will guarantee posterior propriety. Figueiredo (2002) proposed to apply RVM using the popular Jeffreys's prior on the hyperparameters. The necessary conditions that we derive show that the choice of Jeffreys's prior also leads to an improper posterior.

Sparse Bayesian learning models also involve classification models. Precise and accurate classification of tumors is of great importance, as it helps oncologists to detect the type of cancer affecting the patient. Such classification is often based on gene expression data. Mallick et al. (2005) proposed a RKHS based Bayesian classification model which makes use of the finite dimensional solution in (3.2) to build models corresponding to both logistic likelihoods as well as support vector machine related likelihoods. They propose to implement their model by using either proper priors or Jeffreys's prior. Proper priors will, of course, lead to a proper posterior, but we prove that the improper prior (i.e., the Jeffreys's prior implemented in their models) leads to an improper posterior.

Computational methods like Markov chain Monte Carlo (MCMC) or maximization of marginal likelihood are often used to estimate the parameters involved in the sparse Bayesian learning model. Such methods are incapable of providing a red flag in case of posterior impropriety, and hence a theoretical study is required to guarantee the existence of a posterior distribution. A nice illustration of incapability of MCMC in detecting posterior impropriety and its consequences can be found in Hobert and Casella (1996) and Athreya and Roy (2014). The article is structured as follows. In Section 3.2, we explain RVM and a related model proposed by Figueiredo (2002) along with their inference method in detail. Further in Section 3.2, we provide necessary and sufficient conditions for the posterior propriety of RVM and show that the sparse Bayesian learning models proposed by Tipping (2001) and Figueiredo (2002) lead to improper posteriors. In Section 3.3, we provide details about the Bayesian classification models proposed by Mallick et al. (2005) and show that the models are improper under the choice of Jeffreys's prior. Some concluding remarks are given in Section 3.4.

3.2 Relevance Vector Machine and Its Impropriety

Let $\{(y_i, x_i), i = 1, 2, \cdots, n\}$ be the training data, where $y_i \in \mathcal{R}$ is the i^{th} observation for the response variable and $x_i \in \mathcal{R}^p$ is the p dimensional covariate vector associated with y_i. Let $y = (y_1, y_2, \cdots, y_n)^T$. Let $\beta = (\beta_0, \beta_1, \cdots, \beta_n)^T$ and K be the $n \times (n+1)$ matrix whose i^{th} row is given by $K_i^T = \big(1, k_\theta(x_i, x_1), k_\theta(x_i, x_2), \cdots, k_\theta(x_i, x_n)\big)$ where $\{k_\theta(x_i, x_j) : i, j = 1, 2, \cdots, n\}$ are the values of the reproducing kernel and θ is a kernel parameter. The relevance vector machine (RVM) proposed by Tipping (2001) is as follows,

$$y|\beta, \sigma^2 \sim N(K\beta, \sigma^2 I), \tag{3.3a}$$

$$\beta|\lambda \sim N(0, D^{-1}) \ \text{ with } D = diag(\lambda_0, \lambda_1, \cdots, \lambda_n), \tag{3.3b}$$

$$\pi(\lambda_i) \propto \lambda_i^{a-1} \exp\{-b\lambda_i\} \ \ \forall i = 0, 1, 2, \cdots, n, \tag{3.3c}$$

$$\pi\left(\frac{1}{\sigma^2}\right) \propto \left(\frac{1}{\sigma^2}\right)^{c-1} \exp\left\{-\frac{d}{\sigma^2}\right\}, \tag{3.3d}$$

where (a, b, c, d) are user defined hyperparameters. Here $\{\sigma^2, \lambda_i : i = 0, 1, 2, \cdots, n\}$ are assumed apriori independent. Also, β and σ^2 are assumed apriori independent. The kernel parameter θ is typically estimated by cross validation in RVM. Let $\lambda = (\lambda_0, \lambda_1, \cdots, \lambda_n)^T$. For $a > 0$ and $b > 0$, $\pi(\lambda_i)$ is a proper Gamma density with parameters a and b for all $i = 0, 1, 2, \cdots, n$. Similarly, for $c > 0$ and $d > 0$, $\pi(1/\sigma^2)$ is a proper Gamma density with parameters c and d. The posterior distribution of $(\beta, 1/\sigma^2, \lambda)$ corresponding to (3.3) is given by,

$$\pi(\beta, 1/\sigma^2, \lambda|y) = \frac{f(y|\beta, \sigma^2)\pi(\beta, 1/\sigma^2, \lambda)}{m(y)}, \tag{3.4}$$

where $f(y|\beta, \sigma^2)$ is the normal density in (3.3a), $\pi(\beta, 1/\sigma^2, \lambda)$ is the prior density of $(\beta, 1/\sigma^2, \lambda)$ defined in (3.3b)-(3.3d) and $m(y)$ is the marginal density defined as,

$$m(y) = \int_{\mathcal{R}_+^{n+1}} \int_{\mathcal{R}_+} \int_{\mathcal{R}^{n+1}} f(y|\beta, \sigma^2)\pi(\beta, 1/\sigma^2, \lambda) d\beta d\frac{1}{\sigma^2} d\lambda,$$

where $\mathcal{R}_+ = (0, \infty)$. The posterior density given in (3.4) is proper if and only if $m(y) < \infty$.

The user defined hyperparameters can be chosen in such a way that the prior distribution imposed on the parameters turns out to be improper, and in such cases the posterior propriety of the model is no longer guaranteed. The following theorems will provide necessary and sufficient conditions for the posterior propriety of RVM, that is $m(y) < \infty$.

Theorem 3.2.1. *If $b = 0$, i.e., if $\pi(\lambda_i) \propto \lambda_i^{a-1}$ $\forall i = 0, 1, \cdots, n$, then $a \in (-1/2, 0)$ is a necessary condition for the posterior propriety of RVM defined in (3.3).*

Proof: See Appendix B.

Theorem 3.2.2. *Suppose $P_K = K(K^TK)^-K^T$ where $(K^TK)^-$ is a generalized inverse of K^TK. Then (i) and (ii) given below are the sufficient conditions for the posterior propriety of RVM defined in (3.3),*

(i) $a > 0$ and $b > 0$ i.e., the prior on λ_i is a proper Gamma distribution.

(ii) $y^T(I - P_K)y + 2d > 0$ and $n > -2c$.

Proof: See Appendix B.

The sufficient conditions above indicate that although improper priors are not allowed on λ, several improper priors on $1/\sigma^2$ will lead to a proper posterior. Some of the popular choices like $\pi(1/\sigma^2) \propto \sigma^2$ and $\pi(1/\sigma^2) \propto 1$ are covered under the sufficient conditions.

In RVM, for given values of the p covariates say x_{new}, the objective is to predict the corresponding response variable say y_{new}. For predicting y_{new}, one can use the posterior predictive density given by,

$$f(y_{new}|y) = \int_{\mathcal{R}_+^{n+1}} \int_{\mathcal{R}_+} \int_{\mathcal{R}^{n+1}} f(y_{new}|\beta, \sigma^2)\ \pi(\beta, 1/\sigma^2, \lambda|y) d\beta\ d\frac{1}{\sigma^2}\ d\lambda, \tag{3.5}$$

where $f(y_{new}|\beta,\sigma^2)$ is the density of $N(K_{new}^T\beta,\sigma^2)$ with $K_{new}^T = \big(1, k_\theta(x_{new},x_1), k_\theta(x_{new},x_2), \cdot\cdot\cdot, k_\theta(x_{new},x_n)\big)$ and $\pi(\beta, 1/\sigma^2, \lambda|y)$ is the posterior density defined in (3.4). Tipping (2001) approximated the posterior predictive density given in (3.5) by,

$$f(y_{new}|y) = \int_{\mathcal{R}^{n+1}} f(y_{new}|\beta,\hat{\sigma}^2) \;\; \pi(\beta|\hat{\lambda},\hat{\sigma}^2,y)d\beta,$$

where

$$(\hat{\lambda},\hat{\sigma}^2) = \underset{\lambda,\sigma^2}{\operatorname{argmax}}\, \pi(\lambda, 1/\sigma^2|y) = \underset{\lambda,\sigma^2}{\operatorname{argmax}}\, f(y|\lambda,\sigma^2), \tag{3.6}$$

where $\pi(\lambda, 1/\sigma^2|y)$ is the marginal posterior density of $1/\sigma^2$ and λ, and

$$f(y|\lambda,\sigma^2) = \int_{\mathcal{R}^{n+1}} f(y|\beta,\sigma^2) \;\; \pi(\beta|\lambda)d\beta.$$

Using (3.3), simple calculations show that,

$$\begin{aligned} \beta|\hat{\lambda},\hat{\sigma}^2,y &\sim N((K^TK+\hat{D}\hat{\sigma}^2)^{-1}K^Ty, (K^TK\hat{\sigma}^{-2}+\hat{D})^{-1}) \\ \implies \quad y_{new}|y &\sim N(K_{new}^T(K^TK+\hat{D}\hat{\sigma}^2)^{-1}K^Ty, K_{new}^T(K^TK\hat{\sigma}^{-2}+\hat{D})^{-1}K_{new}+\hat{\sigma}^2). \end{aligned}$$

The mean of the above posterior predictive distribution is reported by Tipping (2001) as the predicted response when the observed covariates are x_{new}. In the above posterior predictive distribution used by Tipping (2001), we also observe that the parameters λ and σ^2 are estimated by maximizing the marginal density $f(y|\lambda,\sigma^2)$, and the prior imposed on them is $\pi(\lambda,\sigma^{-2}) \propto 1$ (Indeed the second equality in Equation 3.6 follows due to the use of this uniform prior.). Thus, the prior chosen is improper and is equivalent to choosing the hyperparameters (a,b,c,d) in RVM, given in (3.3) to be $(1,0,1,0)$. This choice of hyperparameters does not satisfy the necessary condition derived in Theorem 3.2.1. Tipping (2001) also mentions that optimizing $f(y|\lambda,\sigma^2)$ can be computationally challenging and hence he proposes to estimate $\log\lambda$ and $\log\sigma^{-2}$ by optimizing $\log f(y|\log\lambda,\log\sigma^{-2})$ and assuming uniform prior on $\log\lambda_i$'s and $\log\sigma^{-2}$, i.e., $\pi(\log\lambda,\log\sigma^{-2}) \propto 1$, which is equivalent to $\pi(\lambda,\sigma^{-2}) \propto \sigma^2 \;\prod_{i=0}^n \lambda_i^{-1}$. Such a prior is also improper and can be assumed by choosing

the hyperparameters (a, b, c, d) to be $(0, 0, 0, 0)$. This choice of hyperparameters also violates the necessary conditions derived in Theorem 3.2.1. Thus, the RVM proposed by Tipping (2001) is based on an improper posterior. Figueiredo (2002) proposed to implement RVM by assuming the Jeffreys's prior on the prior variance parameters of β, i.e., $\pi(1/\lambda_i) \propto \lambda_i \ \ \forall\, i$ which is equivalent to $\pi(\lambda_i) \propto 1/\lambda_i \ \ \forall\, i$. As mentioned before, this improper prior violates the necessary conditions derived in Theorem 3.2.1. Hence the model proposed by Figueiredo (2002) also leads to an improper posterior. Thus, the necessary and sufficient conditions derived in Theorem 3.2.1 and Theorem 3.2.2 will be useful for past researchers to check if their choice of hyperparameters in RVM leads to proper posterior.

3.3 Sparse Bayesian Classification Model and Its Impropriety

Let y be an n dimensional vector containing the observed response variables $\{y_i\}_{i=1}^n$ such that $y_i \in \{0, 1\}\ \forall i$, and let z be an n dimensional vector of latent variables that connect the response variables to the covariates. The Bayesian classification model based on reproducing kernels proposed by Mallick et al. (2005) is as follows,

$$
\begin{aligned}
f(y|z) &\propto \exp\left\{ -\sum_{i=1}^{n} l(y_i, z_i) \right\} \\
z|\beta, \sigma^2, \theta &\sim N(K\beta, \sigma^2 I) \\
\beta|\lambda, \sigma^2 &\sim N(0, \sigma^2 D^{-1}) \ \text{ with } D = diag(\lambda_0, \lambda_1, \cdots, \lambda_n) \\
\pi(\lambda_i) &\propto \lambda_i^{a-1} \exp\{-b\lambda_i\} \ \ \forall i = 1, 2, \cdots, n \\
\sigma^2 &\sim IG(c, d) \\
\theta &\sim U(u_1, u_2)
\end{aligned}
\tag{3.7}
$$

where $y = (y_1, y_2, \cdots, y_n)^T$, $z = (z_1, z_2, \cdots, z_n)^T$, $l(\cdot, \cdot)$ is a loss function, $\beta = (\beta_0, \cdots, \beta_n)^T$, K is the $n \times (n+1)$ matrix whose i^{th} row is given by $K_i^T = \big(1, k_\theta(x_i, x_1), k_\theta(x_i, x_2), \cdots, k_\theta(x_i, x_n)\big)$ where $\{k_\theta(x_i, x_j) : i, j = 1, 2, \cdots, n\}$ are the values of the reproducing kernel, θ is the

parameter in the reproducing kernel, $\lambda = (\lambda_0, \lambda_1, \cdots, \lambda_n)^T$ with λ_0 fixed at a small number and (a, b, c, d, u_1, u_2) are user defined hyperparameters. For $X \sim IG(c, d)$, the density of the random variable X is taken to be, $f(x) \propto x^{-a-1}e^{-b/x}\ I(x > 0)$ and $U(u_1, u_2)$ denotes the uniform distribution on the interval (u_1, u_2). For $a > 0$ and $b > 0$, $\pi(\lambda_i)$ is a proper Gamma density with parameters a and b. The parameters λ_i's, σ^2 and θ are assumed apriori independent. In the case of Jeffreys's prior, the prior is assumed on λ_0 as well i.e., $\pi(\lambda) \propto \prod_{i=0}^{n} \lambda_i^{-1}$.

The above model proposed by Mallick et al. (2005) is quite general in nature, since it encompasses popular models like the logistic model and the support vector machine (SVM) model. Mallick et al. (2005) recommend that the above model be implemented using a Jeffreys's prior on λ and a proper prior on σ^2 or by putting proper priors on λ and σ^2. The following proposition shows that putting a Jeffreys's prior on λ leads to an improper posterior.

Proposition 3.3.1. *If the Jeffreys's prior is assumed on λ in the sparse Bayesian classification model given in (3.7), i.e.,*

$$\pi(\lambda) \propto \prod_{i=0}^{n} \lambda_i^{-1},$$

then the posterior density of the parameters and latent variables of interest, $\pi(\beta, z, \sigma^2, \lambda, \theta|y)$ is improper.

Proof: See Appendix B.

Given values of the p covariates say x_{new}, sparse Bayesian classification model in (3.7) is used to predict the class in which y_{new} belongs to. Since the response variable is binary, i.e., $y_{new} \in \{0, 1\}$, the posterior predictive probability is given by

$$P(y_{new} = 1|y) = \int_{\Omega} P(y_{new} = 1|y, x_{new}, \omega)\pi(\omega|y)\ d\omega,$$

where $\Omega = \mathcal{R}^{n+1} \times \mathcal{R}^n \times \mathcal{R}_+ \times \mathcal{R}_+^n \times (u_1, u_2)$, $\omega = (\beta, z, \sigma^2, \lambda, \theta)$, $\pi(\omega|y)$ is the posterior density of the parameters and latent variables of interest. Since the posterior distribution of parameters and latent variables of interest is not known in closed form, Mallick et al. (2005) construct an MCMC sampler to draw observations from it and use those observations to produce a Monte Carlo estimate of $P(y_{new} = 1|y)$. If the Monte Carlo estimate is greater than 0.5, then y_{new} is predicted to be 1 else 0.

MCMC samplers are incapable of providing a red flag when the posterior distribution is improper. In fact, Hobert and Casella (1996) show that the MCMC draws from an improper posterior distribution seem perfectly reasonable. Empirical diagnostic tools that are commonly employed to check if the MCMC sampler has converged are also incapable of detecting posterior impropriety. Further, such empirical diagnostic tools are vulnerable even if the posterior distribution exists (Dixit and Roy (2017)). Recently, Athreya and Roy (2014) prove that the usual sample average estimators of the posterior means of the parameters will converge to zero with probability 1 if the MCMC chain corresponds to an improper posterior distribution.

3.4 Conclusion

In this article we prove that the RVM with hyperparameters values introduced in Tipping (2001), a Bayesian classification model proposed by Mallick et al. (2005) using Jeffreys's prior and a RVM model proposed by Figueiredo (2002) lead to improper posterior distributions. These three models fall under the category of sparse Bayesian learning models. Among the three, RVM proposed by Tipping (2001) is very popular and has been cited more than 5000 times till date while the Bayesian classification model proposed by Mallick et al. (2005) has been proposed for sensitive tasks like classification of tumors. In order to conduct valid Bayesian analysis using RVM, one can use the necessary and sufficient conditions

for posterior propriety of RVM provided in this article. The Bayesian classification model proposed by Mallick et al. (2005) has been proposed to be implemented with either proper priors or Jeffreys's prior. Since the Jeffreys's prior leads to an improper posterior, it must be implemented with proper priors. Thus, this article provides crucial theoretical developments for sparse Bayesian learning models.

3.5 Appendix A: Some Useful Lemmas and Definition

Lemma 3.5.1.

(a) For RVM given in (3.3),

$$f(y|\lambda,\sigma^2) = \frac{\sigma}{(2\pi)^{n/2}}\ |D|^{1/2}\ |K^TK + D\sigma^2|^{-1/2}\ \exp\left\{-\frac{1}{2}y^T\left(\sigma^2 I + KD^{-1}K^T\right)^{-1}y\right\}.$$

(b) For Bayesian classification model in (3.7),

$$f(z|\lambda,\sigma^2,\theta) = \frac{\sigma^{-n}}{(2\pi)^{n/2}}\ |D|^{1/2}\ |K^TK + D|^{-1/2}\ \exp\left\{-\frac{1}{2\sigma^2}z^T\left(I + KD^{-1}K^T\right)^{-1}z\right\},$$

where $f(z|\lambda,\sigma^2,\theta) = \int_{\mathcal{R}^{n+1}} f(z|\beta,\sigma^2,\theta)\pi(\beta|\sigma^2,\lambda)d\beta$.

Proof:

(a) Note that,

$$\begin{aligned}
f(y|\lambda,\sigma^2) &= \int_{-\infty}^{\infty} f(y|\beta,\sigma^2)\ \pi(\beta|\lambda)\ d\beta \\
&= (2\pi)^{-n-1/2}\ \sigma^{-n}\ |D|^{1/2}\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2\sigma^2}\left((y-K\beta)^T(y-K\beta)+\beta^TD\sigma^2\beta\right)\right\}d\beta \\
&= (2\pi)^{-n-1/2}\ \sigma^{-n}\ |D|^{1/2}\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2}\left(\beta^T(K^TK\sigma^{-2}+D)\beta - 2\frac{y^TK}{\sigma^2}\beta + \frac{y^Ty}{\sigma^2}\right)\right\}d\beta \\
&= (2\pi)^{-n/2}\ \sigma^{-n}\ |D|^{1/2}|K^TK\sigma^{-2}+D|^{-1/2}\exp\left\{-\frac{1}{2}\left(\frac{y^Ty}{\sigma^2} - \frac{y^TK}{\sigma^2}(K^TK+D\sigma^2)^{-1}K^Ty\right)\right\} \\
&= \frac{\sigma}{(2\pi)^{n/2}}\ |D|^{1/2}\ |K^TK + D\sigma^2|^{-1/2}\ \exp\left\{-\frac{1}{2\sigma^2}\left(y^T(I-K(K^TK+D\sigma^2)^{-1}K^T)y\right)\right\}.
\end{aligned}$$

Using Schur complement,

$$\left(\sigma^2 I + KD^{-1}K^T\right)^{-1} = \frac{1}{\sigma^2}(I - K(K^TK + D\sigma^2)^{-1}K^T), \tag{3.8}$$

we get,

$$f(y|\lambda,\sigma^2) = \frac{\sigma}{(2\pi)^{n/2}}\ |D|^{1/2}\ |K^TK + D\sigma^2|^{-1/2}\ \exp\left\{ -\frac{1}{2}y^T\left(\sigma^2 I + KD^{-1}K^T\right)^{-1}y\right\}.$$

(b) Note that,

$$\begin{aligned}
f(z|\lambda,\sigma^2,\theta) &= \int_{-\infty}^{\infty} f(z|\beta,\sigma^2,\theta)\ \pi(\beta|\lambda,\sigma^2)\ d\beta \\
&= (2\pi)^{-n-1/2}\ \sigma^{-2n-1}\ |D|^{1/2}\int_{-\infty}^{\infty}\exp\left\{ -\frac{1}{2\sigma^2}\left((z-K\beta)^T(z-K\beta) + \beta^T D\beta\right)\right\}d\beta \\
&= (2\pi)^{-n-1/2}\ \sigma^{-2n-1}\ |D|^{1/2}\int_{-\infty}^{\infty}\exp\left\{ -\frac{1}{2\sigma^2}\left(\beta^T(K^TK+D)\beta - 2z^TK\beta + z^Tz\right)\right\}d\beta \\
&= (2\pi)^{-n/2}\ \sigma^{-2n-1}\ |D|^{1/2}\ |(K^TK+D)^{-1}\sigma^2|^{1/2}\exp\left\{ -\frac{1}{2\sigma^2}\left(z^Tz - z^TK(K^TK+D)^{-1}K^Tz\right)\right\} \\
&= \frac{\sigma^{-n}}{(2\pi)^{n/2}}\ |D|^{1/2}\ |K^TK+D|^{-1/2}\exp\left\{ -\frac{1}{2\sigma^2}\left(z^T(I - K(K^TK+D)^{-1}K^T)z\right)\right\}.
\end{aligned}$$

As in the proof of part (a), using Schur complement,

$$\left(I + KD^{-1}K^T\right)^{-1} = I - K(K^TK+D)^{-1}K^T,$$

we get,

$$f(z|\lambda,\sigma^2,\theta) = \frac{\sigma^{-n}}{(2\pi)^{n/2}}\ |D|^{1/2}\ |K^TK+D|^{-1/2}\ \exp\left\{ -\frac{1}{2\sigma^2}z^T\left(I + KD^{-1}K^T\right)^{-1}z\right\}.$$

Definition 3.5.1. *Let $r = (r_1, r_2, \cdots, r_n)^T \in \mathcal{R}^n$ and $s = (s_1, s_2, \cdots, s_n)^T \in \mathcal{R}^n$ be any two n dimensional vectors. A real valued function f defined on $\mathcal{R}^n$ is said to be non decreasing in each of its arguments if $r << s$ i.e., $r_i \le s_i\ \forall i \implies f(r) \le f(s)$.*

Lemma 3.5.2. *Let $P_K = K(K^TK)^-K^T$, where $(K^TK)^-$ is a generalized inverse of K^TK. Let $f_1(\lambda_0^{-1}, \lambda_1^{-1}, \cdots, \lambda_n^{-1}) = \exp\left\{ -\frac{1}{2}y^T\left(\sigma^2 I + KD^{-1}K^T\right)^{-1}y\right\}$. Then,*

$$\exp\left\{ -\frac{1}{2\sigma^2}y^Ty\right\} \leq f_1(\lambda_0^{-1}, \lambda_1^{-1}, \cdots, \lambda_n^{-1}) \leq \exp\left\{ -\frac{1}{2\sigma^2}y^T(I - P_K)y\right\}.$$

Proof:

Differentiating f_1 with respect to λ_i^{-1} we get,

$$\frac{\partial f_1}{\partial \lambda_i^{-1}} = \exp\left\{-\frac{1}{2}y^T\left(\sigma^2 I+KD^{-1}K^T\right)^{-1}y\right\}\frac{1}{2}y^T\left(\sigma^2 I+KD^{-1}K^T\right)^{-1}\left(KE_iK^T\right)\left(\sigma^2 I+KD^{-1}K^T\right)^{-1}y,$$

where E_i is a $(n+1)\times(n+1)$ matrix with 1 in the i^{th} diagonal and 0 everywhere else. Since KE_iK^T is positive semidefinite, we get,

$$\frac{\partial f_1}{\partial \lambda_i^{-1}} \geq 0 \quad \forall\ i.$$

$$\implies f_1 \text{ is a non decreasing function of } (\lambda_0^{-1}, \lambda_1^{-1}, \cdots, \lambda_n^{-1}).$$

Also,

$$\lim_{(\lambda_0^{-1}, \lambda_1^{-1}, \cdots, \lambda_n^{-1})\to 0} f_1(\lambda_0^{-1}, \lambda_1^{-1}, \cdots, \lambda_n^{-1}) = \exp\left\{ -\frac{1}{2\sigma^2}y^Ty\right\},$$

where $(\lambda_0^{-1}, \lambda_1^{-1}, \cdots, \lambda_n^{-1}) \to a$ means $\lambda_i^{-1} \to a$ for all $i = 0, 1, \cdots, n$.

Let $\lambda_{min} = min\{\lambda_0, \lambda_1, \cdots, \lambda_n\}$. Then,

$$(K^TK + D\sigma^2)^{-1} \leq (K^TK + \lambda_{min}\sigma^2\ I)^{-1}. \tag{3.9}$$

Using (3.8) and (3.9), we have,

$$f_1(\lambda_0^{-1}, \lambda_1^{-1}, \cdots, \lambda_n^{-1}) \leq \exp\left\{ -\frac{1}{2\sigma^2}\left(y^T(I - K(K^TK + \lambda_{min}\sigma^2\ I_{n+1})^{-1}K^T)y\right)\right\}.$$

By Lemma 1 of Hobert and Casella (1996), we have,

$$\lim_{(\lambda_0^{-1}, \lambda_1^{-1}, \cdots, \lambda_n^{-1})\to \infty} f_1(\lambda_0^{-1}, \lambda_1^{-1}, \cdots, \lambda_n^{-1}) \leq \exp\left\{ -\frac{1}{2\sigma^2}y^T(I - P_K)y\right\}.$$

Thus,

$$\exp\left\{-\frac{1}{2\sigma^2}y^Ty\right\} \leq f_1(\lambda_0^{-1},\lambda_1^{-1},\cdots,\lambda_n^{-1}) \leq \exp\left\{-\frac{1}{2\sigma^2}y^T(I-P_K)y\right\}.$$

The first inequality given above also follows from the fact that $\sigma^2 + KD^{-1}K^T \geq \sigma^2 I$. A similar argument could be used to prove the second inequality if K^TK was non singular.

Lemma 3.5.3. *Consider the following integral,*

$$\int_{\mathcal{R}_+} \frac{t^{-(a+1)}}{(k+t)^{1/2}}dt \tag{3.10}$$

where k and a are constants. The above integral is finite iff $a \in (-1/2, 0)$. In that case, the value of the integral is $ck^{-(a+1/2)}$, where c is some other constant.

Proof:

Considering the transformation $t = k\tan^2\theta$, the integral in (3.10) becomes,

$$2\ k^{-(a+1/2)} \int_0^{\pi/2} (\sec^2\theta - 1)^{-(a+1)} \tan\theta \sec\theta\ d\theta.$$

Letting $z = \sec\theta$, the above integral becomes,

$$2\ k^{-(a+1/2)} \int_1^{\infty} (z^2-1)^{-(a+1)}\ dz.$$

The above integral is finite iff $a \in (-1/2, 0)$, thus proving the first part. Provided the above integral is some finite constant say $c/2$, the value of the integral given in (3.10) becomes $ck^{-(a+1/2)}$. Thus proving the second part of the lemma.

3.6 Appendix B: Proof of Theorems

Proof of Theorem 3.2.1

For RVM defined in (3.3),

$$f(y|\sigma^2) = \int_{\mathcal{R}_+^{n+1}} f(y|\lambda, \sigma^2)\pi(\lambda)d\lambda.$$

Using Lemma 3.5.1 part (a) and substituting $b = 0$ in (3.3), we get,

$$f(y|\sigma^2) = \int_{\mathcal{R}_+^{n+1}} \frac{\sigma}{(2\pi)^{n/2}} |D|^{1/2} |K^TK + D\sigma^2|^{-1/2} \exp\left\{-\frac{1}{2}y^T\left(\sigma^2 I + KD^{-1}K^T\right)^{-1}y\right\} \prod_{i=0}^{n} \lambda_i^{a-1} d\lambda.$$

Let $e_1, e_2, \cdots, e_{n+1}$ be the $n+1$ eigenvalues of K^TK where $e_{max} = max\{e_1, e_2, \cdots, e_{n+1}\}$. Then, $K^TK + D\sigma^2 \leq e_{max}\ I + D\sigma^2$. Hence we get,

$$|K^TK + D\sigma^2|^{-1/2} \geq \prod_{i=0}^{n} \left(\lambda_i\sigma^2 + e_{max}\right)^{-1/2}. \tag{3.11}$$

Using Lemma 3.5.2 and (3.11), we get,

$$f(y|\sigma^2) \geq \frac{\sigma}{(2\pi)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}y^Ty\right\} \prod_{i=0}^{n} \int_{\mathcal{R}_+} \left[\sigma^2 + \frac{e_{max}}{\lambda_i}\right]^{-1/2} \lambda_i^{a-1} d\lambda_i.$$

Letting $t = 1/\lambda_i$, we get,

$$f(y|\sigma^2) \geq \frac{\sigma}{(2\pi)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}y^Ty\right\} \left[\frac{1}{e_{max}^{1/2}} \int_{\mathcal{R}_+} \frac{t^{-(a+1)}}{\left(\frac{\sigma^2}{e_{max}} + t\right)^{1/2}} dt\right]^{n+1}.$$

Using Lemma 3.5.3, the above integral is finite iff $a \in (-1/2, 0)$. Thus proving the necessary condition for the posterior propriety of RVM.

Proof of Theorem 3.2.2

For RVM defined in (3.3),

$$f(y|\sigma^2) = \int_{\mathcal{R}_+^{n+1}} f(y|\lambda, \sigma^2)\pi(\lambda)d\lambda.$$

Using Lemma 3.5.1 part (a) we get,

$$f(y|\sigma^2) = \int_{\mathcal{R}_+^{n+1}} \frac{\sigma}{(2\pi)^{n/2}} \, |D|^{1/2} \, |K^TK + D\sigma^2|^{-1/2} \, \exp\left\{ -\frac{1}{2} y^T \left(\sigma^2 I + KD^{-1}K^T\right)^{-1} y \right\} \pi(\lambda) d\lambda.$$

Since $K^TK + D\sigma^2 \geq D\sigma^2$, we get,

$$|K^TK + D\sigma^2|^{-1/2} \leq \prod_{i=0}^{n} \left(\lambda_i \sigma^2\right)^{-1/2}. \tag{3.12}$$

Using Lemma 3.5.2 and (3.12), we get,

$$f(y|\sigma^2) \leq \frac{1}{(2\pi)^{n/2}} \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left\{ -\frac{1}{2\sigma^2} y^T (I - P_K) y \right\} \int_{\mathcal{R}_+^{n+1}} \pi(\lambda) \, d\lambda. \tag{3.13}$$

As mentioned before, for $a > 0$ and $b > 0$, $\pi(\lambda)$ is a product of $n+1$ proper Gamma densities with parameters a and b. Hence, for $a > 0$ and $b > 0$, the value of the integral in (3.13) will be 1. Further,

$$\begin{aligned} m(y) &= \int_{\mathcal{R}_+} f(y|\sigma^2) \pi\left(\frac{1}{\sigma^2}\right) d\frac{1}{\sigma^2} \\ &\leq \frac{1}{(2\pi)^{n/2}} \int_{\mathcal{R}_+} \left(\frac{1}{\sigma^2}\right)^{n/2+c-1} \exp\left\{ -\frac{1}{\sigma^2} \left(\frac{y^T(I - P_K)y}{2} + d \right) \right\} d\frac{1}{\sigma^2}. \end{aligned}$$

The integral above will be finite if $y^T(I - P_K)y + 2d > 0$ and $n > -2c$, thus proving the sufficient conditions for posterior propriety of RVM.

Proof of Proposition 3.3.1

For Bayesian classification model given in (3.7),

$$f(z|\sigma^2, \theta) = \int_{\mathcal{R}_+^{n+1}} f(z|\lambda, \sigma^2, \theta) \pi(\lambda) d\lambda.$$

Using Lemma 3.5.1 part (b), we get,

$$f(z|\sigma^2, \theta) = \int_{\mathcal{R}_+^{n+1}} \frac{\sigma^{-n}}{(2\pi)^{n/2}} \, |D|^{1/2} \, |K^TK + D|^{-1/2} \, \exp\left\{ -\frac{1}{2\sigma^2} z^T \left(I + KD^{-1}K^T\right)^{-1} z \right\} \pi(\lambda) d\lambda.$$

Since, $I + KD^{-1}K^T \geq I$, we get,

$$\exp\left\{ -\frac{1}{2\sigma^2} z^T \left(I + KD^{-1}K^T\right)^{-1} z \right\} \geq \exp\left\{ -\frac{1}{2\sigma^2} z^T z \right\}. \tag{3.14}$$

Further, $K^TK + D \leq e_{max}\ I + D$ where e_{max} is the largest eigenvalue of K^TK, hence we get,

$$|K^TK + D|^{-1/2} \geq \prod_{i=0}^{n} \left(\lambda_i + e_{max}\right)^{-1/2}. \tag{3.15}$$

Using (3.14) and (3.15), we get,

$$f(z|\sigma^2, \theta) \geq \frac{\sigma^{-n}}{(2\pi)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2} z^T z \right\} \prod_{i=0}^{n} \int_{\mathcal{R}_+} \left[1 + \frac{e_{max}}{\lambda_i}\right]^{-1/2} \lambda_i^{-1} d\lambda_i.$$

Letting $t = 1/\lambda_i$, we get,

$$f(z|\sigma^2, \theta) \geq \frac{\sigma^{-n}}{(2\pi)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2} z^T z \right\} \left[\frac{1}{e_{max}^{1/2}} \int_{\mathcal{R}_+} \frac{t^{-1}}{\left(\frac{1}{e_{max}} + t\right)^{1/2}} dt \right]^{n+1}.$$

From Lemma 3.5.3, the above integral is equal to ∞, thus proving Proposition 3.3.1.

Bibliography

Athreya, K. B. and Roy, V. (2014). Monte carlo methods for improper target distributions. *Electronic Journal of Statistics*, 8(2):2664–2692.

Berlinet, A. and Thomas-Agnan, C. (2011). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer US.

Bishop, C. M. and Tipping, M. E. (2000). Variational Relevance Vector Machines. *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, UAI'00:46–53.

Dixit, A. and Roy, V. (2017). MCMC diagnostics for higher dimensions using Kullback Leibler divergence. *Journal of Statistical Computation and Simulation*, 87(13):2622–2638.

Figueiredo, M. (2002). Adaptive sparseness using Jeffreys prior. *Advances in neural information processing systems*, 15:697–704.

Hobert, J. P. and Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association*, 91(436):1461–1473.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Jeffreys, H. (1961). *Theory of Probability and Inference*. Cambridge Univ. Press, London.

Mallick, B. K., Ghosh, D., and Ghosh, M. (2005). Bayesian classification of tumours by using gene expression data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):219–234.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 58:267–288.

Tipping, M. E. (2000). The Relevance Vector Machine. *Advances in Neural Information Processing Systems*, 12:652–658.

Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.*, 1:211–244.

Wahba, G. (1990). *Spline models for observational data*. Society for Industrial and Applied Mathematics.

CHAPTER 4. ANALYZING RELEVANCE VECTOR MACHINES USING A SINGLE PENALTY APPROACH

A paper under preparation

Anand Dixit and Vivekananda Roy

Abstract

Relevance vector machine (RVM) is a popular sparse Bayesian learning model typically used for prediction. Recently it has been shown that improper priors assumed on multiple penalty parameters in RVM may lead to an improper posterior. Currently in the literature, the sufficient conditions for posterior propriety of RVM do not allow improper priors over the multiple penalty parameters. In this article, we propose a single penalty relevance vector machine (SPRVM) model in which multiple penalty parameters are replaced by a single penalty and analyze it using a semi Bayesian approach. The necessary and sufficient conditions for posterior propriety of SPRVM are more liberal than those of RVM and allow for several improper priors over the penalty parameter. Additionally, we also prove the geometric ergodicity of the Gibbs sampler used to analyze the SPRVM model and hence can estimate the asymptotic standard errors associated with the Monte Carlo estimate of the means of the posterior predictive distribution. Such a Monte Carlo standard error cannot be computed in the case of RVM, since the rate of convergence of the Gibbs sampler used to analyze RVM is not known. The predictive performance of RVM and SPRVM is compared by analyzing a genetic dataset, gas dataset and a cookie dataset.

4.1 Introduction

Suppose we are given a training dataset consisting of n observations and p predictors. Let $\{(y_i, x_i) : i = 1, 2, \cdots, n\}$ denote the training dataset where $y_i \in \mathcal{R}$ is the i^{th} observation of the response variable and $x_i \in \mathcal{R}^p$ is the p dimensional covariate vector associated with y_i. For such a dataset, often the objective is to come up with a function h, such that the response variable y_i can be expressed as $y_i = h(x_i) + \epsilon_i \; \forall \; i = 1, 2, \cdots, n$ where $h : \mathcal{R}^p \to \mathcal{R}$ and $\{\epsilon_i\}_{i=1}^n$ are the corresponding errors. Many times, for a previously unobserved p dimensional covariate vector, the function h is utilized to predict its associated response variable. If p is small, then the function h can be estimated using the nonparametric approach of a Nadaraya-Watson type estimator. In this approach, the errors are assumed to be uncorrelated, have a zero mean and a constant variance. For higher dimensions, kernel density estimation might not work well, and hence Nadaraya Watson type estimators are not recommended when p is large. Thus, in cases where p is large but smaller than n, one can use the ordinary least squares (OLS) method to estimate the function h. In OLS, h is estimated from a class of linear models by minimizing the quadratic loss function.

In recent years, there is a plethora of datasets wherein p is far greater than n. Such datasets are often referred to as high dimensional datasets. Examples of these can be found in the field of genetics, nutrition, chemical engineering etc. In such cases, the methods described before are no longer applicable. A possible solution in such cases is to use the least absolute shrinkage and selection operator (LASSO) proposed by Tibshirani (1996) that estimates the function h from a class of linear models by minimizing the quadratic loss function with respect to an L_1 constraint. Another option is to utilize the ridge estimator proposed by Hoerl and Kennard (1970) that is similar to LASSO, but uses an L_2 constraint instead of an L_1 constraint. There are other variants of LASSO and ridge estimators proposed in the literature. Such methods are commonly referred to as penalized regression methods.

In recent years, the Bayesian version of such methods have gained popularity as they allow researchers to account for prior information in their analysis. Parameter estimation in the Bayesian version is often carried out using Markov chain Monte Carlo (MCMC) samplers. For example, the Bayesian LASSO proposed by Park and Casella (2008) can be analyzed using a Gibbs sampler. In traditional as well as Bayesian versions, a drawback of these penalized regression methods is that the function h is restricted among the class of linear models.

Thus, if we wish to explore a more general class of models, a common strategy is to take a reproducing kernel Hilbert space (RKHS) approach to estimate the function h. Such an estimate of the function h was found by Wahba (1990) by solving the Tikhonov regularization over RKHS. This RKHS based solution allows us to reduce the complexity of the model matrix from p to n dimensions. This pleasing property of the RKHS based solution was utilized by Tipping (2001) to propose the relevance vector machine (RVM) (see also Tipping (2000) and Bishop and Tipping (2000)).

RVM is a hierarchical Bayesian model in which the finite dimensional solution found by Wahba (1990) was utilized as the mean structure of the data model. Further, Tipping (2001) assumed an improper prior over its hyperparameters. Assuming improper priors is fine as long as the posterior propriety has been established. Recently, Dixit and Roy (2018) provide necessary and sufficient conditions for posterior propriety of RVM and prove that improper priors assumed by Tipping (2001) lead to an improper posterior distribution. Thus, in order to conduct valid Bayesian analysis, one needs to either use proper priors or improper priors that satisfy the sufficient conditions.

In the past, Fokou et al. (2011) have attempted to implement RVM using conjugate proper priors over its hyperparameters. In that case, the full conditional distribution of the parameters involved in RVM are well known distributions which are easy to simulate from and hence can be utilized to construct an RVM Gibbs sampler. Further, for a previously

unobserved p dimensional covariate vector, the response variable can be predicted by utilizing the RVM Gibbs sampler iterations to produce a Monte Carlo estimate of the mean of the posterior predictive distribution. A Monte Carlo estimate should ideally be accompanied by a valid standard error estimate, so that the researcher is aware about the uncertainty associated with the estimate. In order to compute Monte Carlo standard errors for Markov chain samples, one needs to establish a Markov chain central limit theorem, which in turn depends on the rate of convergence of the Markov chain (see Jones and Hobert (2001) and Roberts and Rosenthal (2004)). Currently in the literature, the rate of convergence of the Gibbs sampler implemented by Fokou et al. (2011) is not known and hence the Markov chain CLT is not guaranteed. Thus, in the case of RVM, one cannot compute the standard errors associated with the Monte Carlo estimate of the mean of the posterior predictive distribution.

RVM involves multiple shrinkage parameters which are also known as penalty parameters. In this article we propose to replace these multiple penalty parameters by a single penalty parameter. We propose to name this new model as single penalty relevance vector machine (SPRVM) and analyze it using a semi Bayesian approach.

In SPRVM, conjugate priors are assumed on a few parameters and others are estimated by maximizing the marginal likelihood. Further, in the case of SPRVM, the posterior predictive distribution is not known in closed form, and a Gibbs sampler is implemented to produce a Monte Carlo estimate of the mean of the posterior predictive distribution. Additionally, we also prove that the Gibbs sampler implemented in the case of SPRVM converges at a geometric rate, and hence the Markov chain central limit theorem is guaranteed. Thus, in the case of SPRVM, asymptotically valid standard error estimates can be attached to a Monte Carlo estimate of the mean of the posterior predictive distribution. This is an advantage of SPRVM over RVM.

The article is structured as follows. In Section 4.2, we provide details about RVM and its associated Gibbs sampler. In Section 4.3, we introduce and provide details about SPRVM.

In Section 4.4, we conduct data analysis on some real life datasets obtained from the field of genetics, nutrition and chemical engineering to compare the predictive performance of RVM and SPRVM. Some concluding remarks are provided in Section 4.5.

4.2 Relevance Vector Machine

Let $y = (y_1, y_2, \cdots, y_n)$ be the vector of standardized responses where $y_i \in \mathcal{R}$ is the i^{th} standardized observation for the response variable. Further, let $x_i \in \mathcal{R}^p$ denote the covariate vector associated with the i^{th} observation. Let K be the $n \times (n+1)$ kernel matrix whose i^{th} row is given by $K_i^T = \big(1, k_{i1}, k_{i2}, \cdots, k_{in}\big)$ where $\{k_{ij} = k_\theta(x_i, x_j) : i, j = 1, 2, \cdots, n\}$ are the values of the reproducing kernel and θ is a kernel parameter that is typically tuned using cross validation. Also, let $\beta = (\beta_0, \beta_1, \cdots, \beta_n)$. Then, the RVM proposed by Tipping (2001) is as follows,

$$y|\beta, \sigma^2 \sim N_n(K\beta, \sigma^2 I), \tag{4.1a}$$

$$\beta|\lambda_0, \lambda_1, \cdots, \lambda_n \sim N_{n+1}(0, D^{-1}) \;\text{ with } D = diag(\lambda_0, \lambda_1, \cdots, \lambda_n), \tag{4.1b}$$

$$\pi(\lambda_i) \propto \lambda_i^{a-1} \exp\{-b\lambda_i\} \;\; \forall i = 0, 1, 2, \cdots, n, \tag{4.1c}$$

$$\pi\left(\frac{1}{\sigma^2}\right) \propto \left(\frac{1}{\sigma^2}\right)^{c-1} \exp\left\{-\frac{d}{\sigma^2}\right\}, \tag{4.1d}$$

where (a, b, c, d) are hyperparameters that are specified by the user. In the above model, Tipping (2001) assumes that $1/\sigma^2$ and $\{\lambda_i\}_{i=0}^n$ are apriori independent. Further, β and $1/\sigma^2$ are also assumed to be apriori independent. The posterior density of the parameters in RVM is as follows,

$$\pi(\beta, 1/\sigma^2, \lambda_0, \lambda_1, \cdots, \lambda_n|y) = \frac{f(y|\beta, \sigma^2)\pi(\beta, 1/\sigma^2, \lambda_0, \lambda_1, \cdots, \lambda_n)}{m(y)}, \tag{4.2}$$

where $f(y|\beta, \sigma^2)$ is the data model given in (4.1a), $\pi(\beta, 1/\sigma^2, \lambda_0, \lambda_1, \cdots, \lambda_n)$ is the joint prior density obtained from (4.1b) - (4.1d) and $m(y)$ is the marginal likelihood which is also known

as the normalizing constant. When the posterior density given in (4.2) is integrated over the entire parametric space, the integral is equal to 1, provided the normalizing constant exists. Therefore, the posterior distribution is proper if and only if $m(y) < \infty$.

In Bayesian analysis, prior information available with the researchers is generally incorporated by choosing the user defined hyperparameters accordingly. In the case of RVM, Tipping (2001) chose to use improper priors. Improper priors are functions that integrate to infinity and hence are not valid probability density functions. The improper prior assumed by Tipping (2001) can be obtained by choosing (a, b, c, d) to be $(0, 0, 0, 0)$. Dixit and Roy (2018) proved that for this choice of user specified hyperparameters, the normalizing constant $m(y)$ is infinity, and hence the RVM implemented by Tipping (2001) is based on improper posterior distribution. Given the posterior impropriety of RVM for the hyperparameters used by Tipping (2001), we choose to implement RVM using priors that satisfy the sufficient conditions for posterior propriety derived by Dixit and Roy (2018). RVM is typically used for predicting the response variable say y_{new} for a previously unobserved p dimensional covariate vector say x_{new}. Such a prediction is often based on the posterior predictive distribution of the model, which is given by,

$$f(y_{new}|y) = \int_{\mathcal{R}^{n+1}\times\mathcal{R}_+^{n+2}} f(y_{new}|\beta,\sigma^2)\ \pi(\beta, 1/\sigma^2, \lambda_0, \lambda_1, \cdots, \lambda_n|y) d\beta\ d\frac{1}{\sigma^2}\ d\lambda_0\ d\lambda_1 \cdots \lambda_n. \quad (4.3)$$

Further, the mean of the above posterior predictive density can be reported as the predicted value associated with x_{new} and is given by,

$$E(y_{new}|y) = K_{new}^T\ \bar{\beta}_R \quad (4.4)$$

where $K_{new}^T = \big(1, k_\theta(x_{new}, x_1), k_\theta(x_{new}, x_2), \cdots, k_\theta(x_{new}, x_n)\big)$ and $\bar{\beta}_R$ is the posterior mean of the parameter β in the case of RVM model.

Although posterior propriety is guaranteed for priors that satisfy the sufficient conditions for posterior propriety derived by Dixit and Roy (2018), the marginal likelihood is still analytically intractable, and hence the posterior density given in (4.2) is not available in closed

form. Therefore, the posterior mean of β is not available in closed form. A Monte Carlo estimate for the posterior mean of β can be obtained by implementing a Gibbs sampler to draw approximate observations from the posterior distribution. The full conditional distributions of $(\beta, 1/\sigma^2, \lambda_0, \lambda_1, \cdots, \lambda_n)$, required to implement the Gibbs sampler are as follows:

$$\begin{aligned}
\beta|\cdot &\sim N_{n+1}\left(\left(K^TK + D\sigma^2\right)^{-1}K^Ty, \left(K^TK\frac{1}{\sigma^2} + D\right)^{-1}\right) \\
\frac{1}{\sigma^2}|\cdot &\sim Gamma\left(\frac{n}{2} + c,\ \frac{1}{2}||y - K\beta||^2 + d\right) \\
\lambda_i|\cdot &\sim Gamma\left(a + \frac{1}{2},\ \frac{\beta_i^2}{2} + b\right)\ \ \forall\ i = 0, 1, 2, \cdots, n.
\end{aligned}$$

Thus, RVM Gibbs sampler given above is a two component fixed scan sampler in which, for every iteration, $(1/\sigma^2, \{\lambda_i\}_{i=0}^n)$ is drawn given β and then β is drawn given the other variables. Thus, for a previously unobserved p dimensional covariate say x_{new}, the predicted response, i.e., the estimate of the mean of the posterior predictive distribution is given by

$$\hat{E}(y_{new}|y) = K_{new}^T\ \hat{\bar{\beta}}_{R,M}, \tag{4.5}$$

where K_{new}^T is as defined previously and $\hat{\bar{\beta}}_{R,M}$ is the estimate of the posterior mean of β, found by calculating the sample mean of the M observation drawn from the posterior density given in (4.2) using the RVM Gibbs sampler.

The choice of M depends on the Monte Carlo standard error associated with the estimate given in (4.5). If the Monte Carlo standard error associated with (4.5) is deemed large, then it can be reduced by choosing a larger M. On the other hand, if the Monte Carlo standard error is small, computing resources can be conserved by choosing a smaller M. But, since the rate of convergence of the above RVM Gibbs sampler is not known, we cannot compute the Monte Carlo standard error associated with the Monte Carlo estimate given in (4.5). Thus, in the case of RVM, there are no guidelines for choosing a suitable M. Additionally, if proper priors are assumed in RVM, it requires the specification of user defined hyperparameters

(a, b, c, d). Specifying these hyperparameters to assume a non-informative proper prior can be challenging. Therefore, Fokou et al. (2011) proposed to reduce the number of hyperparameter by assuming an extended hierarchical prior structure. The sufficient conditions for posterior propriety of RVM derived by Dixit and Roy (2018) allow for impropriety over $1/\sigma^2$ but not over $\{\lambda_i\}_{i=0}^n$. Hence in the data analysis section of this article, for $1/\sigma^2$ we assume an improper prior, $\pi(1/\sigma^2) \propto \sigma^2$ which can be obtained by choosing $c = d = 0$, and in the case of $\{\lambda_i\}_{i=0}^n$, for the sake of implementation, we choose $a = 0.001$ and $b = 0.01$ which yields a proper Gamma prior with a mean of 0.1 and a variance of 10.

4.3 Single Penalty Relevance Vector Machine

The prior assumed by Tipping (2001) was improper and hence was looked upon to be non-informative and hyperparameter free. But since it leads to an improper posterior distribution, one cannot implement RVM using that improper prior. In this section we will replace multiple penalty parameters with a single penalty parameter and simplify the prior structure to propose single penalty relevance vector machine (SPRVM).

Let $\{(y_i, x_i) : i = 1, 2, \cdots, n\}$ be the training data containing standardized responses and their corresponding covariate vectors, β be the vector of coefficient parameters and K be the $n \times (n + 1)$ kernel matrix, where y_i, x_i, β and K are as defined previously in Section 4.2. Then we propose SPRVM as follows,

$$y|\beta, \xi \sim N(K\beta, \xi^{-1}I), \tag{4.6a}$$

$$\beta|\lambda \sim N(0, \lambda^{-1}I), \tag{4.6b}$$

$$\pi(\lambda) \propto \lambda^{a-1} \exp\{-b\lambda\} \tag{4.6c}$$

where (a, b) are user specified hyperparameters. If a Gamma prior is assumed on ξ, then a Gibbs sampler can be implemented. Such an MCMC sampler does not work well in practice

since the traceplot for the ξ parameter reveals mixing issues. Therefore, for SPRVM, we do not assume any prior over ξ. In SPRVM, although we assume the intercept and the remaining coefficient parameters to have the same prior variance, in the data analysis section of this article we observed that the posterior density estimate of λ was heavily concentrated over small values of the parametric space $\mathcal{R}_+$, and hence one can argue that the resulting prior over β is a diffuse normal prior. For SPRVM, the posterior density of parameters (β, λ), indexed by ξ, is as follows,

$$\pi(\beta, \lambda|y, \xi) = \frac{f(y|\beta, \xi)\pi(\beta, \lambda)}{m_\xi(y)}, \tag{4.7}$$

where $f(y|\beta, \xi)$ is the data model given in (4.6a), $\pi(\beta, \lambda)$ is the joint prior density following from (4.6b)-(4.6c) and $m_\xi(y)$ is the marginal likelihood which is given by,

$$m_\xi(y) = \int_{\mathcal{R}^{n+1}\times\mathcal{R}_+} f(y|\beta, \xi)\pi(\beta, \lambda)\, d\beta\, d\lambda. \tag{4.8}$$

As mentioned previously in Section 4.2, the posterior density given in (4.7) is proper if and only if the marginal likelihood exists, i.e., if $m_\xi(y) < \infty$. For SPRVM, the necessary conditions for the posterior propriety are as follows.

Theorem 4.3.1. *If $b = 0$ i.e., if $\pi(\lambda) \propto \lambda^{a-1}$, then $a \in (-(n+1)/2, 0)$, is a necessary condition for the posterior propriety of SPRVM defined in (4.6).*

Proof: See Appendix B.

The improper priors that do not satisfy the above necessary conditions will lead to an improper posterior. To identify improper priors that will lead to a proper posterior, we need to derive sufficient conditions. Now, the posterior density given in (4.7) is analytically intractable, and therefore, to draw inference from it, one can construct an MCMC sampler. Since the full conditional distributions of (β, λ) are known, we can construct a Gibbs sampler.

The conditionals required for the implementation of the SPRVM Gibbs sampler are as follows,

$$\beta|\cdot \sim N_{n+1}\Big(\big(K^TK+\lambda\xi^{-1}I\big)^{-1}K^Ty, \big(K^TK\xi+\lambda I\big)^{-1}\Big) \tag{4.9a}$$

$$\lambda|\cdot \sim Gamma\Big(\frac{n+1}{2}+a,\ \frac{\beta^T\beta}{2}+b\Big). \tag{4.9b}$$

Let $\{(\beta^{(j)},\lambda^{(j)})\}_{j=0}^{\infty}$ be the fixed scan two component Markov chain associated with the SPRVM Gibbs sampler. Such a Gibbs sampler is geometrically ergodic if there exists a positive real valued function G and a constant $\rho\in[0,1)$ such that,

$$||P^t\big((\beta_0,\lambda_0),\cdot\big)-\Pi(\cdot|y)||_{TV}\leq G(\beta_0,\lambda_0)\rho^t \quad \forall\ \ t=1,2,\cdots \tag{4.10}$$

where $||\cdot||_{TV}$ denotes the total variation norm, $P^t\big((\beta_0,\lambda_0),\cdot\big)$ denotes the probability distribution of the SPRVM Markov chain started at (β_0,λ_0) after t steps and $\Pi(\cdot|y)$ is the probability measure corresponding to the posterior density given in (4.7). If the geometric ergodicity of the SPRVM Gibbs sampler is established, then under finite moments, a central limit theorem is guaranteed for the posterior mean estimates of (β,λ) computed using the SPRVM Gibbs sampler draws (see Roberts and Rosenthal (1997)). The geometric ergodicity of SPRVM Gibbs sampler defined in (4.10) is proved in the following theorem.

Theorem 4.3.2. *The SPRVM Gibbs sampler* $\{(\beta^{(j)},\lambda^{(j)})\}_{j=0}^{\infty}$ *is geometrically ergodic if conditions (i), (ii) and (iii) given below are satisfied.*

(i) Either $b>0$ *or* $a<b=0$.

(ii) There exists $s\in(0,1]$ *such that,*

$$\frac{\Gamma\Big(\frac{n+1}{2}+a-s\Big)}{\Gamma\Big(\frac{n+1}{2}+a\Big)}\ <\ 2^s.$$

(iii) The kernel matrix K defined earlier in Section 4.2 is such that,

$$\frac{k_{ij}}{k_{jj}} \neq 1 \text{ and } k_{jj} \neq 0 \ \forall i, j = 1, 2, \cdots, n \text{ and } i \neq j.$$

Proof: See Appendix B.

Remark 4.3.1. *Taking $s = 1$, condition (ii) of Theorem 4.3.2 holds for $a > \frac{-(n-2)}{2}$.*

Remark 4.3.2. *The following are some examples of reproducing kernels typically used in sparse Bayesian learning models.*

- ***Gaussian kernel:***

$$k_{ij} = k_\theta(x_i, x_j) = \exp\left\{-\frac{||x_i - x_j||^2}{\theta^2}\right\} \ \forall \ i, j = 1, 2, \cdots, n,$$

where $\theta \in \mathcal{R}_+$ and $||\cdot||$ denotes the Euclidean norm.

- ***Laplace kernel:***

$$k_{ij} = k_\theta(x_i, x_j) = \exp\left\{-\frac{||x_i - x_j||}{\theta}\right\} \ \forall \ i, j = 1, 2, \cdots, n,$$

where $\theta \in \mathcal{R}_+$.

- ***Polynomial kernel:***

$$k_{ij} = k_\theta(x_i, x_j) = (1 + x_i^T x_j)^\theta \ \forall \ i, j = 1, 2, \cdots, n,$$

where $\theta \in \mathcal{N}$.

Note that for each of the above three kernels, the condition (iii) of Theorem 4.3.2 will be satisfied if $x_i \neq x_j \ \forall \ i, j = 1, 2, \cdots, n$ and $i \neq j$.

Since the conditions for geometric ergodicity are sufficient for posterior propriety, a large class of improper priors guarantee posterior propriety for SPRVM. There is also a significant

overlap in the necessary and sufficient conditions for posterior propriety. The necessary and sufficient conditions for posterior propriety of RVM derived by Dixit and Roy (2018) do not have any overlap in them. In fact, the sufficient conditions in Dixit and Roy (2018) do not allow for any prior impropriety in multiple penalty parameters of RVM. Given the sufficient conditions for posterior propriety of SPRVM, we propose to assume the following improper prior on the penalty parameter λ,

$$\pi(\lambda) \propto \frac{1}{\lambda^2}. \tag{4.11}$$

From Remark 4.3.1, for $n \geq 5$, the above improper prior satisfies the sufficient condition for posterior propriety of SPRVM. Thus, the above improper prior allows SPRVM to have a non-informative prior structure without the difficulty of specifying any hyperparameters and also leads to a proper posterior as long as $n \geq 5$. Thus, SPRVM is able to achieve the objective of specifying a non-informative improper prior which leads to a proper posterior.

As mentioned earlier, we do not assume any prior over ξ. For a given value of ξ, if the marginal likelihood $m_\xi(y)$ takes a small value, then it indicates that the chosen model is not a good fit to the data y. Therefore a good estimate of ξ, which is denoted by $\hat{\xi}$ is the one for which the marginal likelihood is maximized. Indeed, $\hat{\xi} = \arg\max m_\xi(y)$ is known as the empirical Bayes estimate of ξ. But, as mentioned previously, marginal likelihood is not available in closed form and hence for estimating $\hat{\xi}$, we consider a computational approach described below.

Since the data model and prior on β are both normal densities, we can analytically integrate out β and simplify the right hand side of (4.8). Thus, after analytically integrating out β, the marginal posterior density of λ is as follows,

$$\pi(\lambda|y,\xi) = \frac{f(y|\lambda,\xi)\pi(\lambda)}{m_\xi(y)}, \tag{4.12}$$

where

$$f(y|\lambda,\xi) = \frac{\xi^{-1/2}}{(2\pi)^{n/2}}\, \lambda^{(n+1)/2}\, |K^TK + \lambda\xi^{-1}I|^{-1/2}\, \exp\left\{-\frac{1}{2}y^T\left(\xi^{-1}I + \lambda^{-1}KK^T\right)^{-1}y\right\}, \tag{4.13}$$

and $m_\xi(y)$ is defined previously in (4.8).

Even after integrating out β, $m_\xi(y)$ given in (4.8) is still analytically intractable. Let $B(\xi, \xi_1) = m_\xi(y)/m_{\xi_1}(y)$ be the ratio of marginal likelihoods wherein the denominator is the marginal likelihood evaluated at a fixed value $\xi = \xi_1$. Thus, maximizing $m_\xi(y)$ over $\mathcal{R}_+$ is same as maximizing $B(\xi, \xi_1)$ over $\mathcal{R}_+$ i.e. $\hat{\xi} = \arg\max m_\xi(y) = \arg\max B(\xi, \xi_1)$. Further, the ratio of marginal likelihoods $B(\xi, \xi_1)$ can be written as,

$$B(\xi, \xi_1) = \frac{m_\xi(y)}{m_{\xi_1}(y)} = \int_{\mathcal{R}_+} \frac{f(y|\lambda, \xi)}{f(y|\lambda, \xi_1)} \, \pi(\lambda|y, \xi_1) \, d\lambda. \tag{4.14}$$

Using the above definition, a simple consistent estimator of $B(\xi, \xi_1)$ is as follows,

$$\frac{1}{\tilde{N}} \sum_{i=1}^{\tilde{N}} \frac{f(y|\lambda^{(i)}, \xi)}{f(y|\lambda^{(i)}, \xi_1)} \xrightarrow{a.s} B(\xi, \xi_1), \tag{4.15}$$

where $\{\lambda^{(i)} : i = 1, 2, \cdots, \tilde{N}\}$ are draws from the posterior density $\pi(\beta, \lambda|y, \xi_1)$ obtained using the Gibbs sampler given in (4.9) (Note that the $\{\lambda^{(j)}\}_{j=0}^{\infty}$ sub-chain has the stationary density $\pi(\lambda|y, \xi_1)$.). The importance sampling estimator given in (4.15) will be unstable if the proposal $\pi(\lambda|y, \xi_1)$ and the target $\pi(\lambda|\xi, y)$ differ greatly. Hence, Doss (2010) proposed an importance sampling estimator in which $\pi(\lambda|y, \xi_1)$ is replaced by a mixture of k posterior distributions $\sum_{i=1}^{k} a_i \, \pi(\lambda|y, \xi_i)$, where $\{a_i : i = 1, 2, \cdots, k\}$ are weights. The k points given by $\{\xi_i : i = 1, 2, \cdots, k\}$ are often referred to as skeleton points. Based on the mixture distribution, the estimator proposed by Doss (2010) is given by,

$$\sum_{j=1}^{k} \sum_{l=1}^{\tilde{N}_j} \frac{f(y|\lambda^{(j;l)}, \xi)}{\sum_{i=1}^{k} \frac{\tilde{N}_i}{r_i} f(y|\lambda^{(j;l)}, \xi_i)} \xrightarrow{a.s} B(\xi, \xi_1) \tag{4.16}$$

as $\tilde{N} = \sum_{j=1}^{k} \tilde{N}_j \to \infty$, $\frac{\tilde{N}_i}{\tilde{N}} \to a_i$, where $\{\lambda^{(j;l)} : l = 1, 2, \cdots, \tilde{N}_j\}$ are $\tilde{N}_j$ MCMC draws from $\pi(\lambda|y, \xi_j) \; \forall \; j = 1, 2, \cdots, k$ and $r_i = \frac{m_{\xi_i}}{m_{\xi_1}} \; \forall \; i = 1, 2, \cdots, k$.

The estimator given in (4.16) depends on $r = (r_1, r_2, \cdots, r_k)$ which are unknown. Therefore, Doss (2010) proposed to replace them by estimates found using the reverse logistic

method proposed by Geyer (1994). Thus, replacing $r = (r_1, r_2, \cdots, r_n)$ by its estimate $\hat{r} = (\hat{r}_1, \hat{r}_2, \cdots, \hat{r}_n)$, the estimate of $B(\xi, \xi_1)$ is given by,

$$\hat{B}(\xi, \xi_1) = \sum_{j=1}^{k} \sum_{l=1}^{\tilde{N}_j} \frac{f(y|\lambda^{(j;l)}, \xi)}{\sum_{i=1}^{k} \frac{\tilde{N}_i}{\hat{r}_i} f(y|\lambda^{(j;l)}, \xi_i)}. \tag{4.17}$$

The above estimator can be maximized over $\mathcal{R}_+$ to estimate $\hat{\xi}$. Thus, an estimate of ξ is,

$$\tilde{\xi} = \underset{\xi \in \mathcal{R}_+}{\arg\max}\ \hat{B}(\xi, \xi_1). \tag{4.18}$$

The procedure of estimating ξ can be summarized in the following steps.

- **Step 1:** For each $j = 1, 2, \cdots, k$, simulate observations $\left\{\lambda^{(j;l)}\right\}_{l=1}^{\tilde{n}_j}$ from $\pi(\lambda|y, \xi_j)$ using the SPRVM Gibbs sampler and use them to estimate r by the reverse logistic method proposed by Geyer (1994). Let $\tilde{n} = \sum_{j=1}^{k} \tilde{n}_j$.

- **Step 2:** Independently of Step 1, again simulate observations $\left\{\lambda^{(j;l)}\right\}_{l=1}^{\tilde{N}_j}$ from $\pi(\lambda|y, \xi_j)$ for each $j = 1, 2, \cdots, k$ using the SPRVM Gibbs sampler and estimate the ratio of the marginal likelihood $\hat{B}(\xi, \xi_1)$ by (4.17) using these $\tilde{N} = \sum_{j=1}^{k} \tilde{N}_j$ observations and $\hat{r}$ computed in Step 1.

- **Step 3:** Find $\tilde{\xi}$ by maximizing $\hat{B}(\xi, \xi_1)$ given in (4.17).

In the second step, $\hat{B}(\xi, \xi_1)$ is estimated for large number of ξ values. Thus, to reduce computational burden, the second step sample sizes $\tilde{N}_i$ can't be too large. Additionally, large sample sizes in step 1 result in better estimate of r. The standard error estimates of $\hat{B}(\xi, \xi_1)$ derived in Roy et al. (2018) and Roy and Evangelou (2018) can be used to choose appropriate sample sizes for the two steps. In the past, the above estimation procedure has been implemented by Roy et al. (2016) for estimating parameters in spatial generalized linear mixed models while Roy and Chakraborty (2017) used it for estimating tuning parameters for Bayesian LASSO and Bayesian elastic net.

In the case of SPRVM, prediction for the response variable say y_{new} for a previously unobserved p dimensional covariate vector say x_{new} is based on posterior predictive distribution, which is given by

$$f(y_{new}|y,\tilde{\xi}) = \int_{\mathcal{R}^{n+1}\times\mathcal{R}_+} f(y_{new}|\beta,\tilde{\xi})\ \pi(\beta,\lambda|y,\tilde{\xi})\ d\beta\ d\lambda. \tag{4.19}$$

As observed in the case of RVM, the estimate of the mean of the above posterior predictive distribution which is reported as the predicted response corresponding to x_{new} is given by

$$\hat{E}(y_{new}|y,\tilde{\xi}) = K_{new}^T\ \hat{\bar{\beta}}_{S,M}, \tag{4.20}$$

where K_{new} is as defined previously in (4.5) and $\hat{\bar{\beta}}_{S,M}$ is the estimate of the posterior mean of β found by $\sum_{j=1}^{M}\beta^{(j)}/M$ where $\beta^{(j)}$'s are samples from $\pi(\beta,\lambda|y,\tilde{\xi})$ using the SPRVM Gibbs sampler given in (4.9).

From Theorem 4.3.2, we know that SPRVM Gibbs sampler converges at a geometric rate. Therefore, using Theorem 4.3.2 and assuming $E[\beta^T\beta|y] < \infty$, the following central limit theorem holds,

$$\sqrt{M}\left(\hat{\bar{\beta}}_{S,M} - \bar{\beta}_S\right) \to N(0,\Sigma) \quad \text{as} \quad M\to\infty, \tag{4.21}$$

where $\bar{\beta}_S$ is the posterior mean of β in the case of SPRVM model and Σ is the asymptotic covariance matrix. If the posterior mean estimate i.e. $\hat{\bar{\beta}}_{S,M}$ is based on M iid observations, then Σ can be easily estimated using sample covariance matrix. But since $\hat{\bar{\beta}}_{S,M}$ is based on M draws from the SPRVM Gibbs sampler, the draws are correlated and hence estimating Σ is challenging. In the case of geometrically ergodic Markov chains, Vats et al. (2018) and Vats et al. (2015) have come up with consistent batch means and spectral variance estimators for Σ which can be implemented using the mcmcse R package contributed by Flegal et al. (2017). In the case of SPRVM, the estimate of the standard error associated with the Monte Carlo estimate in (4.20) is given by,

$$\widehat{SE}(K_{new}^T\ \hat{\bar{\beta}}_{S,M}) = \sqrt{K_{new}^T(\hat{\Sigma}/M)K_{new}}, \tag{4.22}$$

where $\hat{\Sigma}$ is a consistent estimator of Σ. Thus, in SPRVM, we can provide a Monte Carlo estimate of the mean of the posterior predictive distribution along with a valid estimate of its standard error.

4.4 Data Analysis

In order to compare the predictive performance of RVM and SPRVM, we implement these two methods on high dimensional datasets in the field of genetics, nutrition and chemical engineering. For each dataset, we split the dataset into training and testing sets. The model is fitted on the training set, and the testing set is utilized to compute the root mean squared prediction error. For both the methods we use the Gaussian kernel and the kernel parameter is tuned by conducting a 5 fold cross validation over the training set. The dataset is split into training and testing set multiple times, and each time the split is done randomly. Thus, at the end we compute the average root mean squared prediction error, which is used to compare the performance of the two methods.

For RVM and SPRVM Gibbs sampler, we run four independent chains using over dispersed starting values for 5000 iterations and assess convergence using potential scale reduction factor (PSRF) proposed by Gelman and Rubin (1992). The PSRF values for all the variables in RVM and SPRVM were close to 1, and thus there was no evidence of non convergence. Therefore, in case of RVM Gibbs sampler and SPRVM Gibbs sampler, the initial 5000 observations are treated as burn-in and observations obtained thereafter are considered to be approximate observations from the posterior distribution. To validate our convergence assessment, we also investigate the corresponding traceplots and observed that the MCMC sampler is fairly stable and there were no signs of non convergence.

In the case of SPRVM, to estimate ξ, we use four skeleton points. Further in the estimation procedure for ξ, at step 1 we choose $\tilde{n}_i = 1800 \; \forall \; i = 1, 2, 3, 4$, and at step 2 we choose

$\tilde{N}_i = 200 \ \forall \ i = 1, 2, 3, 4$. Lastly, for both RVM and SPRVM, in order to draw observations from the posterior predictive distribution, the corresponding Gibbs samplers were run for 10000 iterations out of which first 5000 were treated as burn-in.

The details of the three datasets are as follows.

- **Gene Dataset:** In order to study the genetics of mice population, an experiment was conducted by Lan et al. (2006). For the experiment, a total of 60 mice were available. Among those 60 mice, 31 were females and 29 were males. From each mouse, genetic information corresponding to 22575 genes was collected. Several physiological phenotypes were also collected. We will attempt to predict the physiological phenotype named stearoyl-CoA desaturase (SCD1) using the genetic and gender information available. Thus, our genetic dataset consists of $n = 60$ observations and $p = 22576$ variables. We randomly split the dataset into training set which consists of 50 observations and a testing set which consists of 10 observations. Such splits are performed a total of 20 times. This dataset was analyzed in the past by Zhang et al. (2009) and Bondell and Reich (2012). It can be accessed at (http://www.ncbi.nlm.nih.gov/geo; accession number GSE3330).

- **Gas dataset:** In recent years, chemical engineers have attempted to obtain the octane number of gasoline samples using near infrared (NIR) spectrum measurements. We will work with the gasoline dataset available in the pls R package. The dataset was collected by Kalivas (1997), and the pls R package was contributed by Mevik et al. (2016). The dataset consists of 60 gasoline samples. For each sample, octane number and NIR spectra measurements from 900 nm to 1700 nm in 2nm intervals are provided in the dataset. Thus, our dataset consists of $n = 60$ observations and $p = 401$ variables. The details of splitting the dataset into training and testing are same as the ones mentioned for the gene dataset.

- **Cookie dataset:** In the field of nutrition, researchers are often interested in finding out the fat content of food items. The ppls R package provides a cookie dataset which consists of data on 72 cookie dough samples. For each sample, fat content and NIR spectra measurements from 1100 nm to 2498 nm at 2 nm intervals are provided in the dataset. The R package ppls was provided by Kraemer et al. (2008), and the dataset was collected by Osborne et al. (1984). This dataset was analyzed in the past by Brown et al. (2001) among others. Among the 72 observations, 2 are outliers which are often excluded from analysis. Thus, the dataset consists of $n = 70$ observations and $p = 700$ variables. Out of 70 observations, 50 are randomly chosen to be included in the training set while the remaining 20 are put in the testing set. Such splits are performed 20 times.

The results of our data analysis are provided in the following table.

Table 4.1: Comparing the predictive performance of RVM and SPRVM using average root mean square prediction error

Method	**Cookie dataset**	**Gas dataset**	**Gene dataset**
RVM	0.2689	0.1627	0.8907
SPRVM	0.2613	0.1640	0.8728

In Table 4.1 we observe that the two methods have similar predictive performance. The advantage of SPRVM over RVM is that, we can provide an asymptotically valid standard error estimate along with the Monte Carlo estimate of the mean of the posterior predictive distribution. To provide an illustration, for the gas dataset, consider an out of sample observation in which $y_{new} = 1.1338$. The Monte Carlo estimate of mean of the posterior predictive distribution for that observation was found to be 1.0603 in the case of RVM and 1.0519 in the case of SPRVM. Further, in the case of SPRVM, using (4.22), the associated Monte Carlo standard error was found to be 0.0034. Thus, in the case of SPRVM, we are able to quantify the uncertainty associated with our Monte Carlo estimate.

4.5 Conclusion

In this article we have proposed to analyze RVM using a single penalty parameter instead of multiple penalty parameters. The single penalty relevance vector machine (SPRVM) model was analyzed using a semi Bayesian approach. In the case of SPRVM, the sufficient conditions for posterior propriety allow for several improper priors over the penalty parameter. Currently in the literature, impropriety is not allowed over any of the penalty parameters in RVM. Additionally, we also prove the geometric ergodicity of the Gibbs sampler used to analyze the SPRVM model, and hence using the Markov chain central limit theorem, we can calculate standard errors associated with the Monte Carlo estimate of the mean of the posterior predictive distribution. Such a measure of uncertainty cannot be computed in the case of RVM since the rate of convergence of the RVM Gibbs sampler is currently not known in the literature. Thus, the single penalty relevance vector machine model proposed in this article has advantages over the relevance vector machine.

4.6 Appendix A: Some Useful Lemmas

Lemma 4.6.1. *Let y be an n dimensional vector, K be a $n \times (n+1)$ matrix and $s > 0$. There exists a finite constant Q depending on y and K such that*

$$\left(y^T K (K^T K + \lambda \xi^{-1} I)^{-2} K^T y \right)^s \leq Q.$$

Proof: By definition,

$$K^T K = \sum_{i=1}^{n} t_i t_i^T$$

where t_i^T is the i^{th} row of the matrix K for all $i = 1, 2, \cdots, n$.

The vector y can be expressed as,

$$y = \sum_{j=1}^{n} b_j e_j$$

where for each j, $b_j \in \mathcal{R}$ and e_j is the j^{th} unit vector with 1 in the j^{th} place and 0 everywhere else, $j = 1, 2, \cdots, n$. Therefore,

$$y^T K(K^T K + \lambda\xi^{-1} I)^{-2} K^T y = \left(\sum_{i=1}^{n} b_i e_i^T K\right)(K^T K + \lambda\xi^{-1} I)^{-2}\left(\sum_{j=1}^{n} b_j K^T e_j\right)$$
$$= \sum_{i=1}^{n}\sum_{j=1}^{n} b_i b_j t_i^T \left(\sum_{k=1}^{n} t_k t_k^T + \lambda\xi^{-1} I\right)^{-2} t_j. \tag{4.23}$$

Using Lemma 3 of Khare and Hobert (2011),

$$t_i^T \left(\sum_{k=1}^{n} t_k t_k^T + \lambda\xi^{-1} I\right)^{-2} t_i \leq Q_i \quad \forall\ i = 1, 2, \cdots, n,$$

where $\{Q_i : i = 1, 2, \cdots, n\}$ are constants that depends on $n, t_1, t_2, \cdots, t_n$.

By Cauchy-Schwartz inequality,

$$\left[t_i^T \left(\sum_{k=1}^{n} t_k t_k^T + \lambda\xi^{-1} I\right)^{-2} t_j\right]^2 \leq \left[t_i^T \left(\sum_{k=1}^{n} t_k t_k^T + \lambda\xi^{-1} I\right)^{-2} t_i\right]\left[t_j^T \left(\sum_{k=1}^{n} t_k t_k^T + \lambda\xi^{-1} I\right)^{-2} t_j\right]$$
$$\leq Q_i\ Q_j \quad \forall\ i, j = 1, 2, \cdots, n. \tag{4.24}$$

Taking square root on both sides of (4.24) we get,

$$t_i^T \left(\sum_{k=1}^{n} t_k t_k^T + \lambda\xi^{-1} I\right)^{-2} t_j \leq \sqrt{Q_i\ Q_j} \quad \forall\ i, j = 1, 2, \cdots, n. \tag{4.25}$$

Using (4.23) and (4.25),

$$\left(y^T K(K^T K + \lambda\xi^{-1} I)^{-2} K^T y\right)^s \leq Q$$

where $Q = \left(\sum_{i=1}^{n}\sum_{j=1}^{n} |b_i\ b_j|\ \sqrt{Q_i Q_j}\right)^s$.

Lemma 4.6.2. *Suppose K is a $n \times (n+1)$ kernel matrix defined previously in Section 4.2 that satisfies condition (iii) of Theorem 4.3.2. Then, K is a full row rank matrix.*

Proof: Let $\alpha_i \in \mathcal{R}$ for all $i = 1, 2, \cdots, n$. We need to show that $\alpha_1 = \alpha_2 = \cdots = \alpha_n = 0$ is the only solution that satisfies the following equations,

$$\sum_{i=1}^{n} \alpha_i = 0 \tag{4.26}$$

$$\sum_{i=1}^{n} \alpha_i k_{ij} = 0 \quad \forall\ j = 1, 2, \cdots, n. \tag{4.27}$$

Using (4.26) we get,

$$\alpha_j = -\sum_{\substack{i=1 \\ i \neq j}}^{n} \alpha_i \quad \forall\ j = 1, 2, \cdots, n. \tag{4.28}$$

Using (4.27) we get,

$$\alpha_j = \frac{-1}{k_{jj}} \sum_{\substack{i=1 \\ i \neq j}}^{n} \alpha_i k_{ij} \quad \forall\ j = 1, 2, \cdots, n. \tag{4.29}$$

Further, using (4.28) and (4.29), we get,

$$\sum_{\substack{i=1 \\ i \neq j}}^{n} \alpha_i \left(1 - \frac{k_{ij}}{k_{jj}}\right) = 0 \quad \forall\ j = 1, 2, \cdots, n. \tag{4.30}$$

Using condition (iii) of Theorem 4.3.2. $\exists\ \gamma_1 \in \mathcal{R} - \{0\}$ and $\gamma_2 \in \mathcal{R} - \{0\}$ such that,

$$\gamma_1 \sum_{\substack{i=1 \\ i \neq j}}^{n} \alpha_i \ \leq\ \sum_{\substack{i=1 \\ i \neq j}}^{n} \alpha_i \left(1 - \frac{k_{ij}}{k_{jj}}\right) \ \leq\ \gamma_2 \sum_{\substack{i=1 \\ i \neq j}}^{n} \alpha_i \quad \forall\ j = 1, 2, \cdots, n. \tag{4.31}$$

- **Case 1:** If $\gamma_1, \gamma_2 > 0$ or $\gamma_1, \gamma_2 < 0$, then from (4.30) and (4.31) we have,

$$\sum_{\substack{i=1 \\ i \neq j}}^{n} \alpha_i = 0 \quad \forall\ j = 1, 2, \cdots, n. \tag{4.32}$$

  Using (4.26) and (4.32), $\alpha_1 = \alpha_2 = \cdots = \alpha_n = 0$ is the only possible solution.

- **Case 2:** If $\gamma_1 > 0$ and $\gamma_2 < 0$, then from (4.30) and (4.31) we have,

$$\sum_{\substack{i=1 \\ i \neq j}}^{n} \alpha_i \leq 0 \quad \forall\ j = 1, 2, \cdots, n. \tag{4.33}$$

  Using (4.26) and (4.33), $\alpha_1 = \alpha_2 = \cdots = \alpha_n = 0$ is the only possible solution.

- **Case 3:** If $\gamma_1 < 0$ and $\gamma_2 > 0$, then from (4.30) and (4.31) we have,

$$\sum_{\substack{i=1 \\ i \neq j}}^{n} \alpha_i \geq 0 \quad \forall \; j = 1, 2, \cdots, n. \tag{4.34}$$

Using (4.26) and (4.34), $\alpha_1 = \alpha_2 = \cdots = \alpha_n = 0$ is the only possible solution.

Thus, combining Case 1, Case 2 and Case 3, K is a full row rank matrix, i.e. $rank(K) = n$.

Lemma 4.6.3. *Suppose K is a kernel matrix as defined in Section 4.2 that satisfies condition (iii) of Theorem 4.3.2 and $s \in (0, 1]$ then,*

$$\left[tr\left((K^T K\xi + \lambda I)^{-1}\right)\right]^s \leq \xi^{-s}\left[tr\left((K^T K)^{+}\right)\right]^s + \lambda^{-s},$$

where $(K^T K)^{+}$ denotes the Moore Penrose inverse of $K^T K$.

Proof: Let $O\Psi O^T$ be the spectral decomposition of $K^T K$ where O is an orthogonal matrix such that its columns $\{o_i\}_{i=1}^{n+1}$ are eigenvectors of $K^T K$ and $\Psi = diag(\psi_1, \psi_2, \cdots, \psi_{n+1})$ is a diagonal matrix whose diagonal elements are eigenvalues of $K^T K$. Then,

$$\left(K^T K\xi + \lambda I\right)^{-1} = O\left(\Psi\xi + \lambda I\right)^{-1} O^T. \tag{4.35}$$

As in Abrahamsen and Hobert (2017), let Ψ^{+} be a $(n+1) \times (n+1)$ diagonal matrix whose i^{th} diagonal element is given by,

$$\psi_i^{+} = \psi_i^{-1}(1 - I_{\{0\}}(\psi_i)) \quad \forall \quad i = 1, 2, \cdots, n+1.$$

Further,

$$\left(\psi_i\xi + \lambda\right)^{-1} \leq \; \xi^{-1}\psi_i^{+} + \lambda^{-1} I_{\{0\}}(\psi_i^{+}) \quad \forall \quad i = 1, 2, \cdots, n+1.$$

So,

$$\left(\Psi\xi + \lambda I\right)^{-1} \leq \; \xi^{-1}\Psi^{+} + \lambda^{-1}(I - P_{\Psi}) \tag{4.36}$$

where P_Ψ is a $(n+1)\times(n+1)$ diagonal matrix whose i^{th} diagonal element is $1 - I_{\{0\}}(\psi_i)$. Using (4.35) and (4.36), we get,

$$\begin{aligned}\left(K^TK\xi + \lambda I\right)^{-1} &\leq \xi^{-1}\ O\Psi^+O^T + \lambda^{-1}\ O(I - P_\Psi)O^T \\ &= \xi^{-1}\ (K^TK)^+ + \lambda^{-1}\ O(I - P_\Psi)O^T. \end{aligned} \tag{4.37}$$

Let $\tilde{O}$ be submatrix of O consisting of columns $\{o_i\}_{i\in\mathcal{A}}$ where $\mathcal{A} = \{i \in \{1, 2, \cdots, n+1\} : \psi_i > 0\}$ then,

$$OP_\Psi O^T = \sum_{i\in\mathcal{A}} o_i o_i^T = \tilde{O}\tilde{O}^T.$$

Further, $\tilde{O}\tilde{O}^T$ is an orthogonal projection onto K^TK since $\{o_i\}_{i\in\mathcal{A}}$ forms an orthogonal basis for the column space of K^TK. Therefore,

$$O(I - P_\Psi)O^T = I - P_{K^TK}, \tag{4.38}$$

where P_{K^TK} denotes orthogonal projection onto column space of K^TK.

Using (4.37) and (4.38) and since $s \in (0, 1]$,

$$\begin{aligned}&\left(K^TK\xi + \lambda I\right)^{-1} \leq \xi^{-1}\ (K^TK)^+ + \lambda^{-1}\ (I - P_{K^TK}) \\ \implies &\left[tr\left((K^TK\xi + \lambda I)^{-1}\right)\right]^s \leq \xi^{-s}\left[tr\left((K^TK)^+\right)\right]^s + \lambda^{-s}\left[tr(I - P_{K^TK})\right]^s. \end{aligned} \tag{4.39}$$

Further, using Lemma 2,

$$\begin{aligned} tr(I - P_{K^TK}) &= tr(I) - tr(P_{K^TK}) \\ &= (n+1) - rank(K) \\ &= 1. \end{aligned} \tag{4.40}$$

Using (4.39) and (4.40), we get,

$$\left[tr\left((K^TK\xi + \lambda I)^{-1}\right)\right]^s \leq \xi^{-s}\left[tr\left((K^TK)^+\right)\right]^s + \lambda^{-s}.$$

Hence proved.

Lemma 4.6.4. *Consider the following integral,*

$$\int_{\mathcal{R}_+} \frac{t^{-(a+1)}}{(g+t)^{(n+1)/2}} dt,$$

where g and a are constants. The above integral is finite iff $a \in (-(n+1)/2, 0)$.

Proof: Suppose $t = g\tan^2\omega$, then the above integral becomes,

$$2g^{-(a+(n+1)/2)} \int_0^{\pi/2} \frac{(\tan^2\omega)^{-(a+1)}}{(sec^2\omega)^{(n+1)/2}} \tan\omega \ \sec^2\omega \ d\omega.$$

Let $z = \sec^2\omega$, then the above integral becomes,

$$g^{-(a+(n+1)/2)} \int_1^{\infty} (z-1)^{-(a+1)} \ z^{-((n+1)/2)} \ dz.$$

The above integral is finite iff $a \in (-(n+1)/2, 0)$. Hence proved.

4.7 Appendix B: Proof of Theorems

Proof of Theorem 4.3.1

From (4.8),

$$m_\xi(y) = \int_{\mathcal{R}_+} f(y|\lambda,\xi)\pi(\lambda) \ d\lambda.$$

Using Lemma 1 part (a) of Dixit and Roy (2018),

$$m_\xi(y) = \int_{\mathcal{R}_+} \frac{\xi^{-1/2}}{(2\pi)^{n/2}} \lambda^{(n+1)/2} |K^TK + \lambda\xi^{-1}I|^{-1/2} \exp\left\{-\frac{1}{2}y^T\left(\xi^{-1}I + \lambda^{-1}KK^T\right)^{-1}y\right\} \lambda^{a-1}d\lambda. \tag{4.41}$$

Let $\psi_1, \psi_2, \cdots, \psi_{n+1}$ be eigenvalues of K^TK where $\psi_{max} = \max\{\psi_1, \psi_2, \cdots, \psi_{n+1}\}$.

Now,

$$K^TK + \lambda\xi^{-1}I \leq (\psi_{max} + \lambda\xi^{-1})I$$

$$|K^TK + \lambda\xi^{-1}I|^{-1/2} \geq (\psi_{max} + \lambda\xi^{-1})^{-(n+1)/2}. \tag{4.42}$$

Also,

$$\xi^{-1}I + \lambda^{-1}KK^T \geq \xi^{-1}I$$

$$\therefore \quad \exp\left\{ -\frac{1}{2}y^T\left(\xi^{-1}I + \lambda^{-1}KK^T\right)^{-1}y \right\} \geq \exp\left\{ -\frac{\xi}{2}y^Ty \right\}. \tag{4.43}$$

Using (4.41), (4.42) and (4.43), we get,

$$m_\xi(y) \geq \frac{\xi^{-1/2}}{(2\pi)^{n/2}} \exp\left\{ -\frac{\xi}{2}y^Ty \right\} \int_{\mathcal{R}_+} \frac{\lambda^{a-1}}{\left(\xi^{-1} + \frac{\psi_{max}}{\lambda}\right)^{\frac{n+1}{2}}}\, d\lambda.$$

Using the transformation $t = 1/\lambda$, the above integral becomes,

$$m_\xi(y) \geq \frac{\xi^{-1/2}}{(2\pi)^{n/2}} \exp\left\{ -\frac{\xi}{2}y^Ty \right\} \left[\frac{1}{\psi_{max}}\right]^{(n+1)/2} \int_{\mathcal{R}_+} \frac{t^{-(a+1)}}{\left(\frac{\xi^{-1}}{\psi_{max}} + t\right)^{\frac{n+1}{2}}} dt.$$

Using Lemma 4.6.4, the above integral is finite iff $a \in (-(n+1)/2, 0)$. Hence proved.

Proof of Theorem 4.3.2

Since, SPRVM Gibbs sampler is a two block Gibbs sampler, the two sub-chains $\{\beta^{(j)}\}_{j=0}^{\infty}$ and $\{\lambda^{(j)}\}_{j=0}^{\infty}$ are themselves Markov chains. Further, the rate of convergence of the three chains $\{\beta^{(j)}, \lambda^{(j)}\}_{j=0}^{\infty}$, $\{\beta^{(j)}\}_{j=0}^{\infty}$ and $\{\lambda^{(j)}\}_{j=0}^{\infty}$ is the same (see Roberts and Rosenthal (2001)). Therefore, if we prove the geometric ergodicity of one of the chains, it holds for all the three chains. We will work with the $\{\lambda^{(j)}\}_{j=0}^{\infty}$ chain. The Markov transition density associated with the $\{\lambda^{(j)}\}_{j=0}^{\infty}$ chain is given by,

$$p_l(\tilde{\lambda}|\lambda) = \int_{\mathcal{R}^{n+1}} \pi(\tilde{\lambda}|\beta, y)\ \pi(\beta|\lambda, y)\ d\beta,$$

where $\pi(\tilde{\lambda}|\beta, y)$ is the density corresponding to the full conditional distribution given in (4.9b) and $\pi(\beta|\lambda, y)$ is the density of the full conditional distribution given in (4.9a).

We define the drift function as follows,

$$v(\tilde{\lambda}) = \tilde{\lambda}^m + \tilde{\lambda}^{-s}, \tag{4.44}$$

where $m \in (0,1)$ is a positive constant that is determined in the proof and $s \in (0,1]$ is a constant that satisfies condition (ii) in Theorem 4.3.2.

Since the above drift function is unbounded off compact sets and $\{\lambda^{(j)}\}_{j=0}^{\infty}$ is a Feller chain, geometric ergodicity of the $\{\lambda^{(j)}\}_{j=0}^{\infty}$ chain is established by proving the following drift condition (see Meyn and Tweedie (1993)),

$$E[v(\tilde{\lambda})|\lambda] = \int_{\mathcal{R}_+} v(\tilde{\lambda})\ p_l(\tilde{\lambda}|\lambda)\ d\tilde{\lambda} \leq L + \rho v(\lambda)$$

where $L > 0$ and $\rho \in (0,1)$ are finite constants.

Note that,

$$E[v(\tilde{\lambda})|\lambda] = E\big[E[v(\tilde{\lambda})|\beta]|\lambda\big]. \tag{4.45}$$

We start with the inner expectation in (4.45). Also, first consider $b > 0$,

$$E[\tilde{\lambda}^m|\beta] = \frac{\Gamma(a+m+\frac{n+1}{2})}{\Gamma(a+\frac{n+1}{2})}\left(\frac{\beta^T\beta}{2}+b\right)^{-m} \leq \frac{\Gamma(a+m+\frac{n+1}{2})}{\Gamma(a+\frac{n+1}{2})}\ b^{-m}. \tag{4.46}$$

Now we consider the outer expectation in (4.45). From (4.46) we get,

$$E[\tilde{\lambda}^m|\lambda] \leq \frac{\Gamma(a+m+\frac{n+1}{2})}{\Gamma(a+\frac{n+1}{2})}\ b^{-m}. \tag{4.47}$$

Next,

$$E[\tilde{\lambda}^{-s}|\beta] = \frac{\Gamma(a-s+\frac{n+1}{2})}{\Gamma(a+\frac{n+1}{2})}\left(\frac{\beta^T\beta}{2}+b\right)^{s} \leq \frac{\Gamma(a-s+\frac{n+1}{2})}{\Gamma(a+\frac{n+1}{2})}\left(\frac{(\beta^T\beta)^s}{2^s}+b^s\right). \tag{4.48}$$

From (4.48), using Lemma 4.6.1 and Lemma 4.6.3 we have,

$$\begin{aligned}
E[\tilde{\lambda}^{-s}|\lambda] &\leq \frac{\Gamma(a-s+\frac{n+1}{2})}{\Gamma(a+\frac{n+1}{2})}\left(\frac{1}{2^s}E[(\beta^T\beta)^s|\lambda]+b^s\right)\\
&\leq \frac{\Gamma(a-s+\frac{n+1}{2})}{\Gamma(a+\frac{n+1}{2})}\left(\frac{1}{2^s}\{E[(\beta^T\beta)|\lambda]\}^s+b^s\right)\\
&\leq \frac{\Gamma(a-s+\frac{n+1}{2})}{\Gamma(a+\frac{n+1}{2})}\left\{\frac{1}{2^s}\left[y^TK(K^TK+\lambda\xi^{-1}I)^{-2}K^Ty+\left(tr\left((K^TK\xi+\lambda I)^{-1}\right)\right)\right]^s+b^s\right\}\\
&\leq \frac{\Gamma(a-s+\frac{n+1}{2})}{\Gamma(a+\frac{n+1}{2})}\frac{1}{2^s}\left[Q+(2b)^s+\xi^{-s}\left(tr\left((K^TK)^+\right)\right)^s+\lambda^{-s}\right]\\
&\leq L_0+\frac{\Gamma(a-s+\frac{n+1}{2})}{\Gamma(a+\frac{n+1}{2})}\frac{1}{2^s}\lambda^{-s}
\end{aligned} \tag{4.49}$$

where $L_0 = \dfrac{\Gamma(a-s+\frac{n+1}{2})}{\Gamma(a+\frac{n+1}{2})}\dfrac{1}{2^s}\left[Q+(2b)^s+\xi^{-s}\left(tr\left((K^TK)^+\right)\right)^s\right]$.

Using (4.44), (4.47) and (4.49), we get,

$$E[v(\tilde{\lambda})|\lambda] \leq L_1 + \rho_0\ v(\lambda)$$

where

$$L_1 = L_0 + \frac{\Gamma(a+m+\frac{n+1}{2})}{\Gamma(a+\frac{n+1}{2})}\ b^{-m} \text{ and } \rho_0 = \frac{\Gamma(a-s+\frac{n+1}{2})}{\Gamma(a+\frac{n+1}{2})}\frac{1}{2^s}$$

are finite constants. Further, using condition (ii) of Theorem 4.3.2, $\rho_0 \in (0,1)$. Thus proving geometric ergodicity of the SPRVM Gibbs sampler for $b > 0$.

Now consider $b = 0$ and $a < 0$. Let Σ_β denote the covariance matrix of $\beta|\lambda, y$ i.e.

$$\Sigma_\beta = (K^TK\xi + \lambda I)^{-1} \quad \Longrightarrow \quad \Sigma_\beta^{-1} = K^TK\xi + \lambda I.$$

As defined in Lemma 4.6.3, let $O\Psi O^T$ be spectral decomposition of K^TK where $\Psi = diag(\psi_1, \psi_2, \cdots, \psi_{n+1})$. Also, let $\psi_{max} = max\{\psi_1, \psi_2, \cdots, \psi_{n+1}\}$. Therefore,

$$\begin{aligned}
&\Sigma_\beta^{-1} \leq (\psi_{max}\ \xi + \lambda)I \\
\Longrightarrow\ &\beta^T\Sigma_\beta^{-1}\beta \leq \beta^T(\psi_{max}\ \xi + \lambda)I\beta \\
\Longrightarrow\ &(\beta^T\Sigma_\beta^{-1}\beta)^{-m} \geq \left[\beta^T(\psi_{max}\ \xi + \lambda)I\beta\right]^{-m}. \qquad (4.50)
\end{aligned}$$

Now, $\beta^T\Sigma_\beta^{-1}\beta|\lambda, y$ has a non central χ^2 distribution with $n+1$ degrees of freedom. Using Lemma 4 of Román and Hobert (2012), for $m \in (0,1)$, we get,

$$E[(\beta^T\Sigma_\beta^{-1}\beta)^{-m}|\lambda] \leq 2^{-m}\ \frac{\Gamma(\frac{n+1}{2}-m)}{\Gamma(\frac{n+1}{2})}. \qquad (4.51)$$

Now, using (4.50) and (4.51),

$$\begin{aligned}
E[(\beta^T\beta)^{-m}|\lambda] &= (\psi_{max}\ \xi + \lambda)^m E\left[\left(\beta^T(\psi_{max}\ \xi + \lambda)I\beta\right)^{-m}\middle|\lambda\right] \\
&\leq (\psi_{max}\ \xi + \lambda)^m E[(\beta^T\Sigma_\beta^{-1}\beta)^{-m}|\lambda] \\
&\leq ((\psi_{max}\ \xi)^m + \lambda^m)\ 2^{-m}\ \frac{\Gamma(\frac{n+1}{2}-m)}{\Gamma(\frac{n+1}{2})}. \qquad (4.52)
\end{aligned}$$

Since,

$$E[\tilde{\lambda}^m|\beta] = \frac{\Gamma(a+m+\frac{n+1}{2})}{\Gamma(a+\frac{n+1}{2})}\frac{1}{2^{-m}}E[(\beta^T\beta)^{-m}|\lambda], \tag{4.53}$$

using (4.52) we have,

$$E[\tilde{\lambda}^m|\lambda] \leq \frac{\Gamma(a+m+\frac{n+1}{2})}{\Gamma(a+\frac{n+1}{2})}\frac{\Gamma(\frac{n+1}{2}-m)}{\Gamma(\frac{n+1}{2})}((\psi_{max}\ \xi)^m + \lambda^m). \tag{4.54}$$

Using (4.44), (4.49) and (4.54),

$$E[v(\tilde{\lambda})|\lambda] \leq \tilde{L_0} + L_1 + \rho_0\lambda^{-s} + \rho_1\lambda^m$$

where ρ_0 is as defined before and

$$\tilde{L_0} = \frac{\Gamma(a-s+\frac{n+1}{2})}{\Gamma(a+\frac{n+1}{2})}\frac{1}{2^s}\left[Q + \xi^{-s}\left(tr\left((K^TK)^+\right)\right)^s\right],$$

$$L_1 = \frac{\Gamma(a+m+\frac{n+1}{2})}{\Gamma(a+\frac{n+1}{2})}\ \frac{\Gamma(\frac{n+1}{2}-m)}{\Gamma(\frac{n+1}{2})}(\psi_{max}\ \xi)^m,$$

$$\rho_1 = \frac{\Gamma(a+m+\frac{n+1}{2})}{\Gamma(a+\frac{n+1}{2})}\frac{\Gamma(\frac{n+1}{2}-m)}{\Gamma(\frac{n+1}{2})}.$$

For $m \in (0,1)\cap(0,-a)$, Román and Hobert (2012) have shown that $\rho_1 < 1$. Let $L^\star = \tilde{L_0} + L_1$ and $\rho\star = \max\{\rho_0, \rho_1\}$. Then for $b = 0$ and $a < 0$,

$$E[v(\tilde{\lambda})|\lambda] \leq L^\star + \rho^\star v(\lambda)$$

where $L^\star$ and $\rho^\star$ are finite constants. Further, $\rho^\star \in (0,1)$ since $\rho_0 \in (0,1)$ and $\rho_1 \in (0,1)$. Thus, we have proved geometric ergodicity of $\{\lambda^{(j)}\}_{j=0}^{\infty}$ for $b = 0$ and $a < 0$. Hence proved.

Bibliography

Abrahamsen, T. and Hobert, J. P. (2017). Convergence analysis of block Gibbs samplers for Bayesian linear mixed models with $p > n$. *Bernoulli*, 23(1):459–478.

Bishop, C. M. and Tipping, M. E. (2000). Variational Relevance Vector Machines. *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, UAI'00:46–53.

Bondell, H. D. and Reich, B. J. (2012). Consistent high-dimensional Bayesian variable selection via penalized credible regions. *Journal of the American Statistical Association*, 107(500):1610–1624.

Brown, P. J., Fearn, T., and Vannucci, M. (2001). Bayesian wavelet regression on curves with application to a spectroscopic calibration problem. *Journal of the American Statistical Association*, 96(454):398–408.

Dixit, A. and Roy, V. (2018). Posterior impropriety of some sparse Bayesian learning models. *submitted*.

Doss, H. (2010). Estimation of large families of Bayes factors from Markov chain output. *Statistica Sinica*, pages 537–560.

Flegal, J., Hughes, J., Vats, D., and Dai, N. (2017). *mcmcse: Monte Carlo Standard Errors for MCMC*. R package version 1.3-2.

Fokou, E., Sun, D., and Goel, P. (2011). Fully Bayesian analysis of the relevance vector machine with an extended hierarchical prior structure. *Statistical Methodology*, 8(1):83 – 96.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–472.

Geyer, C. J. (1994). Estimating normalizing constants and reweighting mixtures. *Technical Report 568, School of Statistics, University of Minnesota*.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Jones, G. L. and Hobert, J. P. (2001). Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Statist. Sci.*, 16(4):312–334.

Kalivas, J. (1997). Two data sets of near infrared spectra. *Chemometrics and Intelligent Laboratory Systems*, 37(2):255 – 259.

Khare, K. and Hobert, J. P. (2011). A spectral analytic comparison of trace-class data augmentation algorithms and their sandwich variants. *The Annals of Statistics*, 39(5):2585–2606.

Kraemer, N., Kraemer, N., Boulesteix, A.-L., and Tutz, G. (2008). Penalized partial least squares with applications to b-spline transformations and functional data. *Chemometrics and Intelligent Laboratory Systems*, 94(60-69).

Lan, H., Chen, M., Flowers, J. B., Yandell, B. S., Stapleton, D. S., Mata, C. M., Mui, E. T., Flowers, M. T., Schueler, K. L., and Manly, K. F. (2006). Combined expression trait correlations and expression quantitative trait locus mapping. *PLoS Genetics*, 2(1):e6.

Mevik, B., Wehrens, R., and Liland, K. H. (2016). *pls: Partial Least Squares and Principal Component Regression*. R package version 2.6-0.

Meyn, S. P. and Tweedie, R. L. (1993). *Markov chains and stochastic stability*. Springer.

Osborne, B. G., Fearn, T., Miller, A. R., and Douglas, S. (1984). Application of near infrared reflectance spectroscopy to the compositional analysis of biscuits and biscuit doughs. *Journal of the Science of Food and Agriculture*, 35(1):99–105.

Park, T. and Casella, G. (2008). The Bayesian LASSO. *Journal of the American Statistical Association*, 103(482):681–686.

Roberts, G. O. and Rosenthal, J. S. (1997). Geometric ergodicity and hybrid Markov chains. *Electron. Commun. Probab.*, 2:13–25.

Roberts, G. O. and Rosenthal, J. S. (2001). Markov chains and de-initializing processes. *Scandinavian Journal of Statistics*, 28(3):489–504.

Roberts, G. O. and Rosenthal, J. S. (2004). General state space Markov chains and MCMC algorithms. *Probab. Surveys*, 1:20–71.

Román, J. C. and Hobert, J. P. (2012). Convergence analysis of the Gibbs sampler for Bayesian general linear mixed models with improper priors. *The Annals of Statistics*, 40(6):2823–2849.

Roy, V. and Chakraborty, S. (2017). Selection of tuning parameters, solution paths and standard errors for Bayesian lassos. *Bayesian Analysis*, 12(3):753–778.

Roy, V. and Evangelou, E. (2018). Selection of proposal distributions for generalized importance sampling estimators. *ArXiv e-prints*.

Roy, V., Evangelou, E., and Zhu, Z. (2016). Efficient estimation and prediction for the Bayesian binary spatial model with flexible link functions. *Biometrics*, 72(1):289–298.

Roy, V., Tan, A., and Flegal, J. (2018). Estimating standard errors for importance sampling estimators with multiple markov chains. *Statistica Sinica*, 28:1079 – 1101.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 58:267–288.

Tipping, M. E. (2000). The Relevance Vector Machine. *Advances in Neural Information Processing Systems*, 12:652–658.

Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.*, 1:211–244.

Vats, D., Flegal, J. M., and Jones, G. L. (2015). Multivariate output analysis for markov chain monte carlo. *arXiv preprint arXiv:1512.07713*.

Vats, D., Flegal, J. M., and Jones, G. L. (2018). Strong consistency of multivariate spectral variance estimators in Markov chain Monte Carlo. *Bernoulli*, 24(3):1860–1909.

Wahba, G. (1990). *Spline models for observational data*. Society for Industrial and Applied Mathematics.

Zhang, D., Lin, Y., and Zhang, M. (2009). Penalized orthogonal-components regression for large p small n data. *Electronic Journal of Statistics*, 3:781–796.

CHAPTER 5. CONCLUSION

In this dissertation we have presented three articles that we hope are a valuable addition to the literature of MCMC diagnostics and sparse Bayesian learning models. In this chapter we list the broad conclusions of the three articles and some areas of improvements that will hopefully lead to future research works.

In the first article we identify drawbacks of existing MCMC diagnostic tools and propose tools based on Kullback Leibler (KL) divergence and smoothing methods to address the observed drawbacks. Since the Tool 1 proposed in that article uses a testing of hypothesis framework, it posses the ability to asses the joint convergence of multiple variables which several existing tools are not able to do so. In the case of non convergence of multiple chains, it is often of interest to identify the reasons behind it. To accomplish this we provide a simple visualization tool that is easy to implement and allows researchers to identify possible reasons for non convergence of multiple chains. In case of multimodal stationary distribution, there is a possibility that MCMC sampler may get stuck in a particular mode. In such a situation, several existing diagnostic tools may be fooled into thinking that stationary distribution is unimodal and hence, may falsely detect convergence. The Tool 2 proposed in the first article incorporates the stationary distribution known uptill the unknown normalizing constant and hence has the ability to detect non convergence in case of multimodal stationary distribution. Currently in the literature there is no theoretical proof showing that the adaptive kernel density estimates based on MCMC samples converges to the stationary density with respect to KL divergence measure and hence is a possible area of improvement for future researchers.

In the second article of this dissertation we provide theoretical developments in sparse Bayesian learning models. We notice that relevance vector machine (RVM) proposed by

Tipping (2001), RVM model proposed by Figueiredo (2002) and a sparse Bayesian classification model based on Jeffreys's prior proposed by Mallick et al. (2005) are models that have an improper prior structure and they simply assumed posterior propriety without providing a theoretical proof of it. We show that all the three models lead to an improper posterior. Further, we also derive necessary and sufficient conditions for posterior propriety of RVM which will be beneficial to RVM users in choosing priors that allow them to conduct valid Bayesian analysis. Given the popularity of RVM (more than 5000 citations till date) and applications of the classification model to sensitive tasks like tumor classification, these theoretical developments are crucial.

In the third article of this dissertation, we provide a single penalty approach to analyzing RVM. Since the rate of convergence of the Gibbs sampler implemented to analyze RVM is not known, it is not possible to calculate asymptotically valid standard error that quantifies the uncertainty associated with Monte Carlo estimate of the mean of the posterior predictive distribution. Such a standard error can be calculated in the case of single penalty relevance vector machine (SPRVM), since we prove the geometric ergodicity of the associated Gibbs sampler. An area of future work can be to see if the Gibbs sampler associated with RVM converges at a geometric rate. This will allow researchers to compare the predictive performance of RVM and SPRVM with respect to the uncertainty associated with its predicted values.

BIBLIOGRAPHY

Figueiredo, M. (2002). Adaptive sparseness using Jeffreys prior. *Advances in neural information processing systems*, 15:697–704.

Hobert, J. P. and Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association*, 91(436):1461–1473.

Mallick, B. K., Ghosh, D., and Ghosh, M. (2005). Bayesian classification of tumours by using gene expression data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):219–234.

Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.*, 1:211–244.

Vats, D., Flegal, J. M., and Jones, G. L. (2015). Multivariate output analysis for markov chain monte carlo. *arXiv preprint arXiv:1512.07713*.

Vats, D., Flegal, J. M., and Jones, G. L. (2018). Strong consistency of multivariate spectral variance estimators in Markov chain Monte Carlo. *Bernoulli*, 24(3):1860–1909.